# Unobtrusive Pedestrian Identification by Leveraging Footstep Sounds with Replay Resistance

*Long Huang and Chen Wang*

*ACM Interact. Mob. Wearable Ubiquitous Technol. 2021*

**Chorom Hamm**

crhamm@mmlab.snu.ac.kr

Aug 3, 2022

# Outline

- Background
- Introduction
- System Design and Details
- Evaluation
- Conclusion

# Pedestrian Identification

- Indoor pedestrian identification is necessary to automate services for smart building
  - Security enhancement system/gate access, patient monitoring, parental control, elderly care, customized environment, and energy saving
- The prior works necessarily require the user's active participation
  - Authentication methods based on secret knowledge or the biometrics
- A camera-based method is one of the solutions, but it has various weaknesses
  - High installation overhead, limitations by view angles and light conditions, and privacy concerns

# Introduction

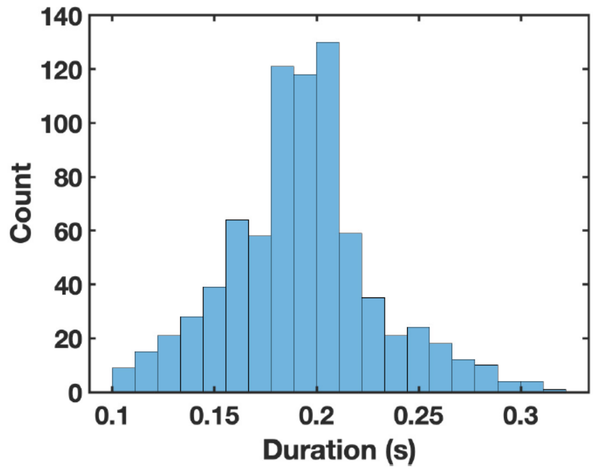
- The paper proposes an unobtrusive pedestrian identification system by passively recognizing the sound of human gait with a low cost
- Voice assistant devices are used to capture multi-dimensional acoustic information without hardware modification
- Footstep sounds can be learned and recognized using a CNN-based algorithm, which tolerates the differences of shoes, floors, and sounds from the left and right foot
- The system is designed and evaluated to prevent replay attacks using the liveness detection method by Differential Time Difference of Arrival (DTDoA)
- It achieves up to 94.9% accuracy in one footstep with various impact factors

# Related Work

- The feature extraction of gait patterns can use a camera, floor sensor, motion sensor, and radio signal, and each has various pros and cons
  - Costs of additional hardware, lack of Line-Of-Sight (LOS), limited sensing ranges…
- Passive acoustic sensing is difficult to obtain sufficient info because the collected sound of steps is too short
  - Active acoustic sensing can be sensed only in a limited range
- Acoustic sensing systems are vulnerable to replay attacks, synthesis attacks, adversarial machine leaning attacks, and ultrasound attacks
  - These attacks should be prevented by detecting the unique liveness using Doppler radars, Time Difference of Arrivals, and magnetic fields

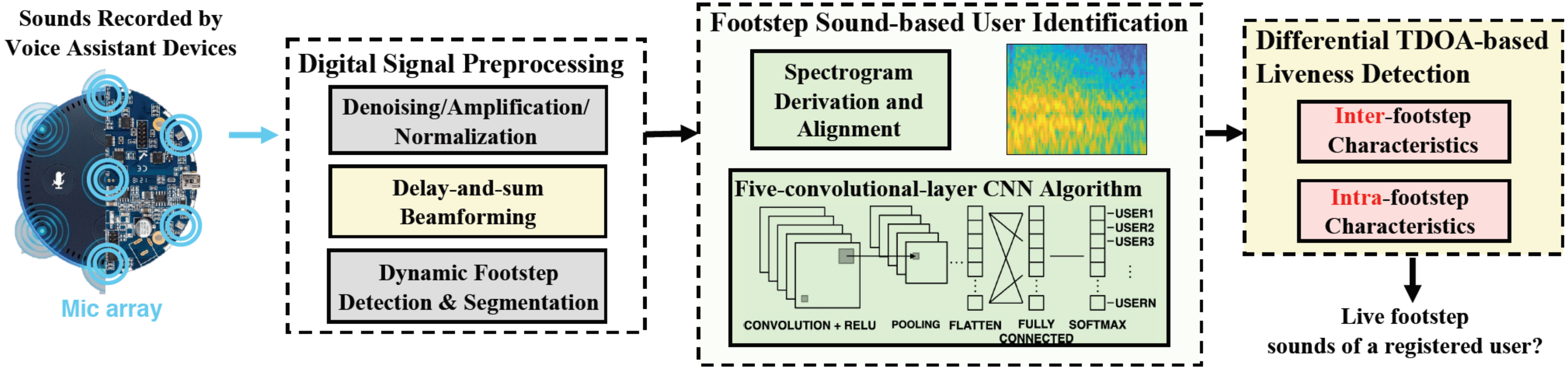
# Footstep Sound Characteristics

- Footstep sounds are related to physiological traits, behavioral characteristics, and shoe and floor types
  - Physiological traits: weights, leg shapes, and foot geometry
  - Behavioral characteristics: the bodyweight shifting from the heel to the sole and from one foot to the other
- The sounds have a low volume and last for a short period
- There are previous studies based on the Mel Frequency Cepstral Coefficient (MFCCs) and machine learning algorithms, but the accuracy is not high



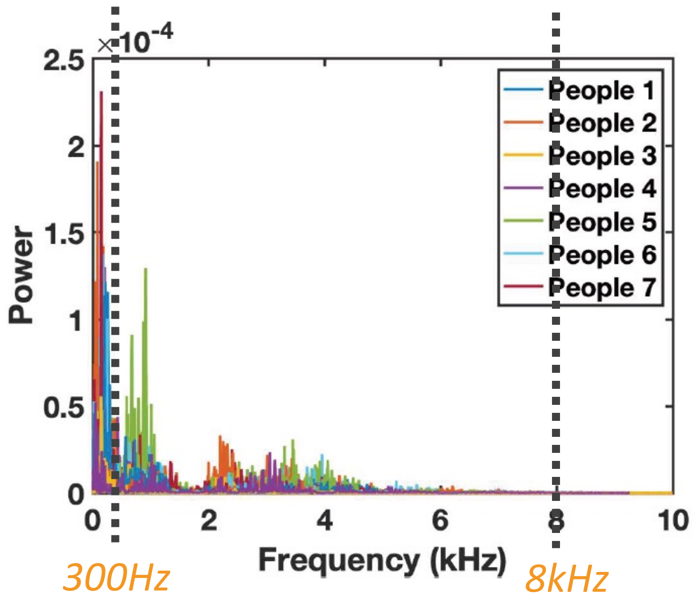
# System Architecture

- The footstep sounds are unobtrusively recorded regardless of walking routes
- The user can be identified by leaning footstep spectrograms based on CNN algorithms, and the TDoA-based liveness detection determines whether it actually belongs to the registered user



# Digital Signal Preprocessing

- The bandpass filter is designed to leave only the signal corresponding to the footstep and remove mechanical vibration noises
- Hampel filter is added to eliminate the outliers
- The delay-and-sum beamforming is applied to improve the SNR of the footstep sound
  - 3dB SNR gains can be achieved using beamforming
- Normalization is essential to focus on the relations of the sound amplitudes and reduce external impacts



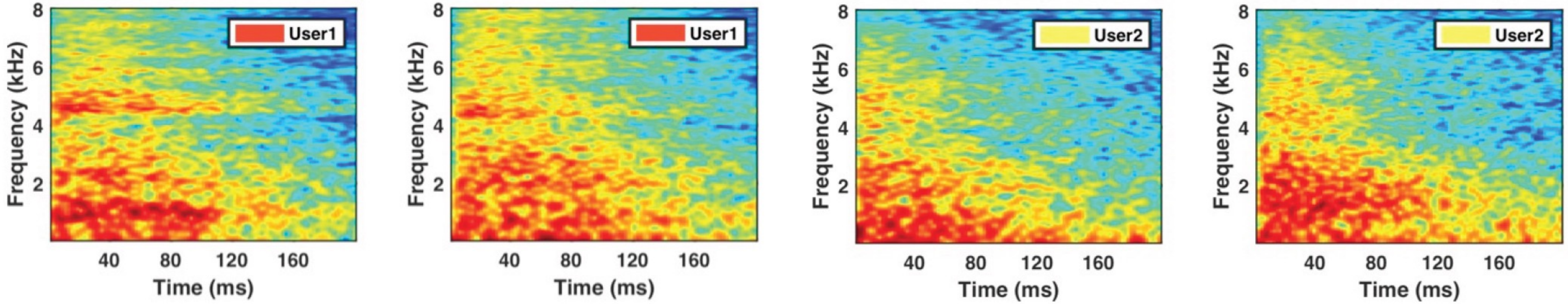

$$bf(k) = \sum_{1}^{n} mic_i(k - tdoa(i,1))$$

# Footstep Detection & Segmentation

- An MFCC-based method is designed using a sliding Hamming window to detect and segment the footstpes
- The system uses the first and second MFCC coefficients for better accuracy
  - Traditionally, step detection proceeded based on short-time energy or moving variance-based methods
- The footstep starting and ending points are determined by the peak and a threshold
- Window size is fixed that is set to be the median time length based on statistics i.g., 200ms

# Footstep Spectrogram Derivation

- Among two phases of gait, the foot-floor contact is focused on user identification, not inter-footstep leg movement
- The sound from the foot-floor contract is translated into spectrums along time
  - Spectrograms show clearly consistent pattern in the same users and distinctive patterns with different users



# CNN-based User Identification

- To classfy a user's footstep sounds with one label regardless of various impacts, the CNN model is used
- Rectified Linear Unit (ReLU) is added to improve training speed
- A 3X3 max-pooling layer makes the feature maps downsampled to reduce computational costs

| Layer | Parameter # | Output Shape | Activation # |
|---|---|---|---|
| Input: Footstep Spectrogram | | (40,98,1) | 3920 |
| Conv2D + RecLineU | 120 | (40,98,12) | 47070 |
| Max Pooling | | (20,49,12) | 11760 |
| Batch Normalization | 24 | (20,49,12) | 11760 |
| Conv2D + RecLineU | 2616 | (20,49,24) | 23520 |
| Max Pooling | | (10,25,24) | 6000 |
| Batch Normalization | 48 | (10,25,24) | 6000 |
| Conv2D + RecLineU | 10416 | (10,25,48) | 12000 |
| Max Pooling | | (5,13,48) | 3120 |
| Batch Normalization | 96 | (5,13,48) | 3120 |
| Conv2D + RecLineU | 20784 | (5,13,48) | 3120 |
| Conv2D + RecLineU | 20784 | (5,13,48) | 3120 |
| Max Pooling | | (5,1,48) | 240 |
| Dropout | | (5,1,48) | 240 |
| Fully Connected + Softmax | 6748 | (28) | 28 |
| Output: Probability Distribution | | (1) | 0 |

- Normalization helps to increase stability of neural network and training speed

# Treat Models

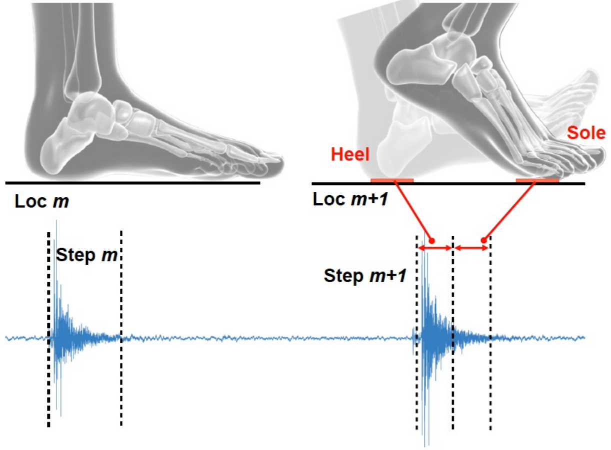
- An attacker can use the sound of footsteps to spoof the identity of the registered user
- *A blind Attack* means the case when there is no intention or when an attacker uses his own gait pattern
- *Human Impersonation Attack* is the case of mimicking the walking behavior of the registered user based on previous experiences
- *Machine-based Impersonation* is an attack using a machine speaker including replay attacks, adversarial examples, and ultrasound attacks
  - Audible or inaudible, comprehensible or not, fixed location or location changes…

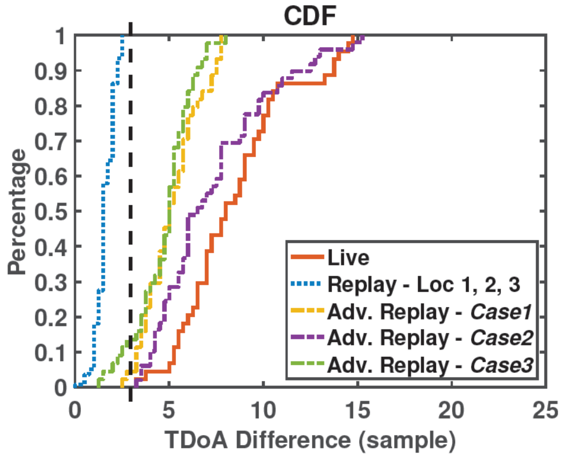
# Footstep Liveness Detection

- Footstep liveness detection method is designed to defend against the machine speaker-based impersonation attacks including replay attacks, adversarial examples, and ultrasound attacks
- Two types of footstep liveness indicators are derived containing the inter-footstep and the intra footstep characteristics
- A supervised-learning method is used to learn two indicators’ statistics from all registered users and set the thresholds for detecting liveness of footsteps
- Advanced machine-speaker impersonation is also prevented by system design
  - It means the case of recording footstep sounds from machine speaker during mobility

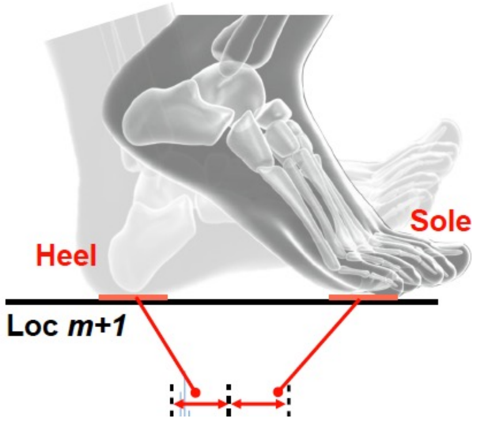
# Inter-footstep Characteristic

- Spatial changes of consecutive footsteps can be a clue to predict whether the footstep is coming from a human or machine-speaker
- It is computed by DTDoA from adjacent footsteps using > 2 microphones
  - DTDoA of machine speaker sounds may show stable or close-to-zero

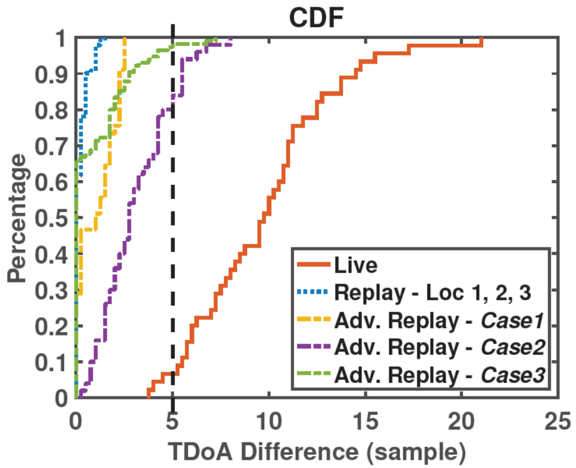


# Intra-footstep Characteristic




- The system should cover 3 cases of advanced replay attacks
  - Case 1. recording only the replayed footstep sounds
  - Case 2. adversary's footstep sound + replayed footstep in the same segment
  - Case 3. detecting adversary's footstep sound in separate footstep segment
- To defend against these attacks, the system uses the unique spatial variations of a single footstep
  - In a footstep, there must be spatial separation: heel striking and foot sole pedaling
  - The intra-footstep characteristic can be calculated by DTDoA of two halves of a footstep segment

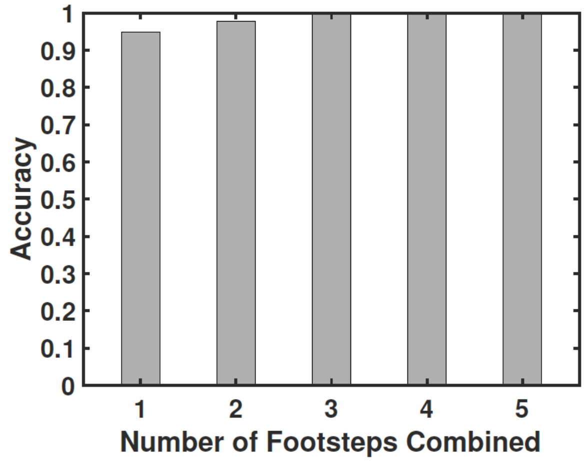


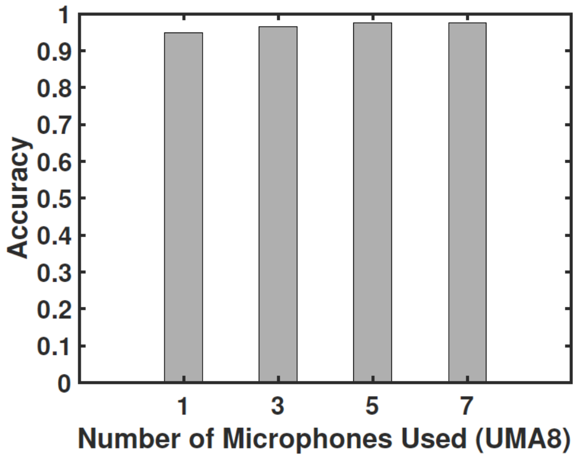
# Performance Evaluation

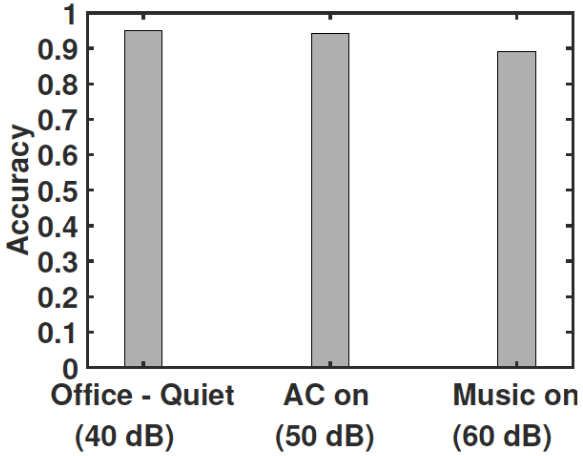
- Voice assistant devices are used for experiments: Samsung Galaxy Note5, Galaxy S8, and UMA8, the microphone array used by Amazon Echo
- The data is collected at different location of voice assistant devices, types of shoes, floor types, walking speeds, levels of ambient noise…
- Evaluation metrics are selected as accuracy, TPR, and TNR
  - Accuracy: correctly identified users over the total users
  - True Positive Rate (TPR): the ratio of correctly classified target users over the total target users
  - True Negative Rate (TNR): how the system prevents attacks and rejects legitimate users

# User Identification Results

- The system achieves 94.9% accuracy based on one microphone of UMA8 using only one footstep
- Identification accuracy comes out to over 97.6% with 5 and 7 microphones
- The accuracy degrades to 94.3% and 89.1% with noise levels of 50dB & 60dB

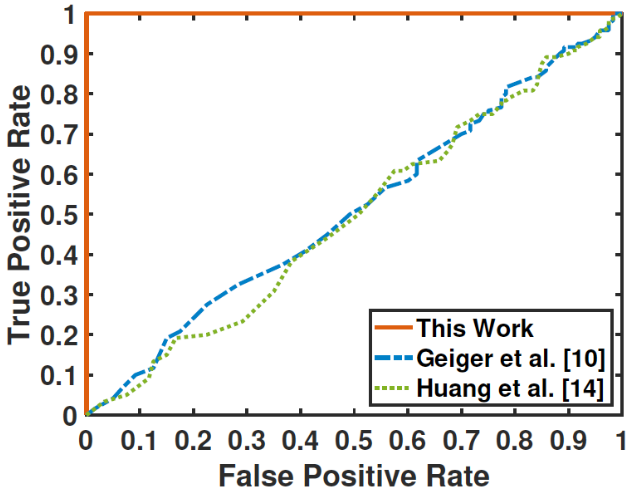


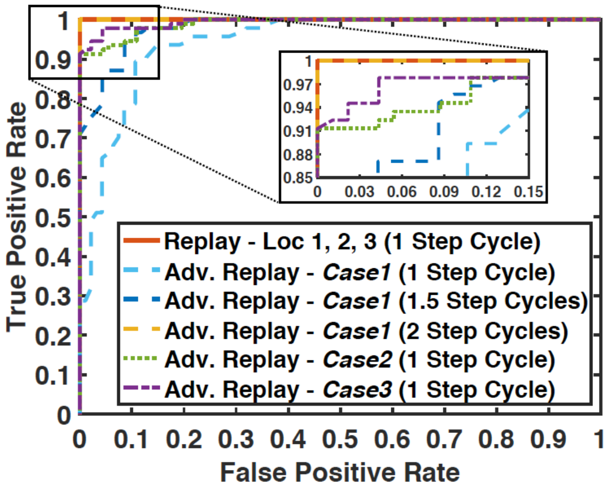



# Performance Under Attacking Scenarios

- Replay attack using a fixed voice assistant device can be prevented with 100% TPR based on one left footstep and the right one
- Advanced replay attacks can be prevented 93.5% TPR in Case2 and 97.8% TPR in Case3 by detecting the unique liveness indicators



# Conclusion

- The paper proposes an unobtrusive pedestrian identification for smart buildings by footstep sound recognition
- It exploits the advanced stereo recording technology of voice assistant devices
- It achieves almost 95% accuracy based on a CNN-based deep learning algorithm using spectrograms of the user’s gait information as inputs
- The beam-forming Is performed to improve the footstep sound SNR
- To prevent replay attacks, it adopts inter-footstep and intra-footstep characteristics as liveness indicators

# Thank you