Audio Fingerprinting for Multi-Device Self-Localization

Tsz-Kin Hon, Lin Wang, Joshua D. Reiss, and Andrea Cavallaro

***Abstract*—We investigate the self-localization problem of an ad-hoc network of randomly distributed and independent devices in an open-space environment with low reverberation but heavy noise (e.g. smartphones recording videos of an outdoor event). Assuming a sufficient number of sound sources, we estimate the distance between a pair of devices from the extreme (minimum and maximum) time difference of arrivals (TDOAs) from the sources to the pair of devices without knowing the time offset. The obtained inter-device distances are then exploited to derive the geometrical configuration of the network. In particular, we propose a robust audio fingerprinting algorithm for noisy recordings and perform landmark matching to construct a histogram of the TDOAs of multiple sources. The extreme TDOAs can be estimated from this histogram. By using audio fingerprinting features, the proposed algorithm works robustly in very noisy environments. Experiments with free-field simulation and open-space recordings prove the effectiveness of the proposed algorithm.**

***Index Terms*—Ad-hoc microphone array, audio fingerprinting, multi-source, self-localization, time difference of arrival (TDOA) estimation.**

I. Introduction

THE diffusion of smartphones has created new opportunities for applications when multiple devices are used to spontaneously capture audio and video of real-world scenes [1]. Device localization is an important task in this context as knowledge of the geometrical configuration of the sensors is necessary in most multi-microphone (e.g. beamforming and sound source localization [2]) and multi-camera (e.g. target tracking with camera networks [3]) signal processing algorithms.

Device localization approaches may use various sensors embedded in smartphones such as GPS, camera and microphone [4]–[6]. While GPS can directly provide physical locations, the accuracy may be unsatisfactory [4]. The distance between smartphones can be calculated via image processing [5]. However, the performance of image-based techniques is confined by the field-of-view of the camera, which requires overlapping views across cameras and known focal lengths. Using acoustic emissions, the inter-distance of two smartphones can be calculated based on the sound time of arrival (TOA) [6]. Sound-based techniques are not limited by orientation and relative position of the smartphones.

Several challenges, such as asynchronous sampling and unknown time offset between devices, arise when localizing (unconnected) devices with sound [1]. Asynchronous sampling can be compensated for in advance with prior knowledge of the smartphones, or using radio signals for synchronizing local clocks [7], [8]. The unknown time offset is mainly due to the unknown processing time of the devices, which causes sending and receiving uncertainties. This problem may be solved by transmitting specially designed acoustic anchor signals (e.g. chirp signal) between devices [6], [9]. However, the active collaboration and interaction between independent devices may not always be feasible. Considering that sound is ubiquitous in real-life scenarios, it would be useful to estimate the inter-device distances using unspecified sounds. However, it is challenging to blindly estimate the time of arrival of the sound reaching each device with unknown time offsets.

Recently, it has been shown that the distance between a pair of devices can be directly computed without knowing the time offsets between the two devices from the time difference of arrivals (TDOAs) of the sound sources located at end-fire positions. End-fire positions are all the points that lie on a line that connects the two devices with the exception of any points between the two devices [10], [11]. The maximum and minimum TDOA pair contains the same distance and time offset information between the two devices, thus making it possible to calculate the inter-device distance by cancelling the time offset. The inter-device distances can be further exploited to derive the geometrical configuration of the whole ad-hoc network. A generalized cross-correlation (GCC)-based algorithm is further proposed to estimate such maximum and minimum TDOAs from multiple sound sources [10], [11], assuming that in each time frame at most one sound source is dominant. This assumption, however, might lead to degraded performance in a noisy environment with multiple simultaneously active sources.

In this paper we focus on sound-based device localization in an outdoor environment where mobile devices such as smartphones capture events. Three features characterize such an acoustic scenario: the reverberation is typically low, the recording is typically noisy, and there are multiple sound sources. Using the same inter-device distance estimation framework and the same assumption on a sufficient number of sound sources and positions as in [10], [11], we propose a novel audio-fingerprinting-based extreme (minimum and maximum) TDOA estimation algorithm. We show that, by increasing time analysis resolution, landmark audio fingerprinting [12] can

Manuscript received November 27, 2014; revised March 31, 2015; accepted May 25, 2015. Date of publication June 05, 2015; date of current version June 17, 2015. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ono.

The authors are with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: tsz.kin.hon@qmul.ac.uk; lin.wang@qmul.ac.uk; joshua.reiss@qmul.ac.uk; a.cavallaro@qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2442417

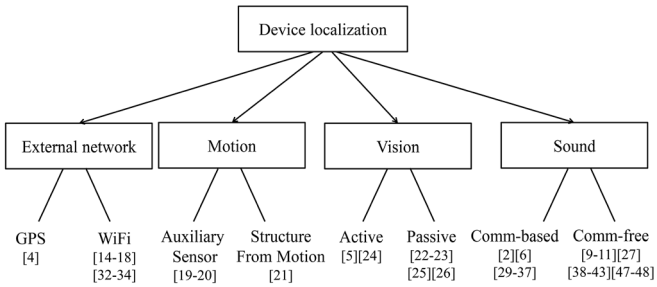


Fig. 1. Device localization methods can be categorized into four classes based on their modalities.

detect the TDOAs of surrounding sound sources captured by two devices. We construct a histogram of the TDOAs from multiple sources by matching the audio landmarks of the recordings from the two devices. We use a metric based on the W-disjoint orthogonality (WDO) [13] to determine the value of the threshold parameters in audio fingerprinting. While landmark-based audio fingerprinting has been widely used in music information retrieval [12] due to its robustness to noise, to the best of our knowledge this is the first time that audio fingerprinting is employed for extreme TDOA estimation.

II. Related Work

Device localization methods can be broadly categorized into four classes (Fig. 1) based on the selected modality: external network, motion, vision, or sound.

External network-based methods depend on external systems, such as satellites and wireless network access points. GPS is commonly used for positioning outdoor devices. GPS localization error ranges from a few meters in an open environment to more than 80 meters in metropolitan areas [4]. Some Wi-Fi-based methods extract the characteristics of the Wi-Fi signals propagating in different environments and construct a fingerprint database for each location of interest [14], [15]. However, accurate estimation is not assured due to volatile radio propagation. Other Wi-Fi-based methods depend on connectivity measurements (hop-count) from anchor points of known positions to mobile devices and are known as range-free localization methods [16]–[18]. Range-free localization methods require a dense and uniform distribution of mobile devices, and can only provide coarse location estimation.

Motion-based methods use the motion of devices to estimate locations. Using auxiliary sensors, such as foot mounted inertial measurement units to measure the acceleration and orientation information, some approaches [19], [20] estimate the position with tracking algorithms, such as Kalman filtering. These approaches suffer from cumulative errors and the localization accuracy drops over time. Another approach constructs the trajectory of the camera and 3D coordinates of a stationary target simultaneously using Structure From Motion [21].

Vision-based methods [22], [23] localize the devices based on their relative distance to a target object, and can be categorized into active and passive approaches. In active approaches, devices need to send reference signals for localization. A projected stripe or spot of light on a stationary object is viewed by a camera, and the distance between camera and object can be determined with known camera focal length and projection angle [24]. In passive approaches, the relative positions between cameras and objects can be determined without sending reference signals. For instance, a network of non-overlapping cameras can be localized using the trajectories of a moving target [25]. Vision-based methods suffer in the presence of motion blur and camera shake, or when the camera focal length is unknown [26].

TABLE I
Summary of the Relevant Methods for Sound-Based Device Localization. (Comm.: Communication)

Comm.	Ref.	Sound type	Method
Yes	[2], [6], [29]–[34]	Calibration sound	TOA-based inter-device distance calculation
	[35]–[37]	Ultrasound	
No	[9], [27], [38]–[43]	Calibration sound	Joint sensor/source localisation based on TOA/TDOAs of multiple sources
	[47], [48]	Ambient sound	Matching diffuse noise coherence
	[10], [11]		Extreme TDOA estimation using GCC
	Proposed		Extreme TDOA estimation using AF

Sound-based methods (Table I) estimate the locations of devices based on the acoustic propagation delays and attenuation [6], [27]. Useful information that can be extracted from the acoustic signals includes received signal strength indication, time of arrival, time difference of arrival and angle of arrival. While the distance estimation accuracy can be as small as a few centimeters, sound-based methods usually need to exchange the timestamps of the local clocks for synchronization [27]. Special hardware is used to tackle time misalignment and to ensure real-time signal sending/receiving [28]. Sound-based methods can be classified based on the communication between independent devices. Communication-based methods need extra collaboration and interaction between devices for synchronization, while communication-free methods utilize external acoustic events either from controlled emissions from external transmitters or from independent ambient sounds. Communication-based methods typically estimate inter-device distance by sending and receiving calibration sounds (e.g. chirp signal) [2], [6], [29]–[37]. Through two-way communication, the internal transmitting/capturing delays of the devices can be naturally cancelled out. In [35]–[37], inaudible ultrasound is used for the communication between devices. In [32]–[34], the locations of the devices are coarsely estimated with WiFi-based methods and then improved with active sound ranging.

Communication-free approaches, known also as passive or self-localization methods, use only external sounds to localize the devices. One can measure the TOAs or TDOAs of the sound sources (from either controlled emissions or from ambient sounds) and then jointly estimate the locations of the sources and sensors [38], [39]. In some ad-hoc configurations, the unknown onset times of the sound sources and the internal delays of the sensors also need to be estimated from the TOA or TDOA measurements [40]–[43]. While various iterative methods have been used to solve this optimization problem, joint estimation of many parameters makes the problem non-convex.

Existing iterative methods are sensitive to the initialization and can get stuck in local minima. In addition to this, sufficient TOA or TDOA measurements are required to make the estimation problem solvable. A mobile beacon is used to send calibration signals to obtain TOA information, whereas radio signals are used to synchronize the clock between recording devices [9], [27].

An alternative for passive localization is to estimate the pairwise distance of the devices from ambient sounds and to recover the relative locations of the devices using a closed-form estimator such as multidimensional scaling [44]–[46]. An approach matches the measured noise coherence to the theoretical model of the sound field for estimating the inter-device distances [47], [48]. This approach is only applicable to relatively small arrays and the assumption of a diffuse noise field is not always met in practical applications. Another approach [10] [11] computes the inter-device distances from the minimum and maximum TDOAs of sound arriving at the ad-hoc network and derives the relative locations of the devices in the network from the obtained inter-device distances. A GCC-based algorithm is proposed to estimate the extreme TDOAs. This approach assumes that the minimum and maximum TDOAs come from the sources at end-fire locations with respect to each pair of devices. With this assumption, the unknown time offset between two asynchronous devices can be cancelled out. This approach can compute the inter-device distance without knowing the time-offset between two devices.

In this paper we use the same framework in [10], [11] as a baseline to develop our method. The assumption of the end-fire source is quite strict but may hold in special acoustic scenarios such as a meeting room where each speaker is located with a laptop, or noisy outdoor environments with a sufficient number of sound sources, like recordings of social events with smartphones.

III. Preliminaries

A. Device Localization via Extreme TDOA

Consider an anechoic environment with an ad-hoc microphone array consisting of N independent devices and unknown number of K sources randomly distributed around the array. Let $\mathbf{R} = [\mathbf{r}_1 \cdots \mathbf{r}_b \cdots \mathbf{r}_K]^T \in \mathbb{R}^{K\times P}$ be the unknown physical locations of the sound sources $s_1, \cdots, s_K$, and $\mathbf{M} = [\mathbf{m}_0 \cdots \mathbf{m}_i \cdots \mathbf{m}_{N-1}]^T \in \mathbb{R}^{N\times P}$ be the unknown locations of the devices with embedded microphone (where T denotes transpose, and P denotes the dimension of the space). The signal recorded at each microphone is denoted as

$$a_i(t) = \sum_{b=1}^{N} g_{ib} s_b(t - t_{ib}), \quad i = 1, \cdots, M \tag{1}$$

where $s_b(\cdot)$, t_{ib} and g_{ib} are the b-th source signal, the propagation time and the attenuation from the b-th source to the i-th microphone, respectively. The inter-device distance $d_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|$ is the Euclidean distance between the locations of a pair of devices. The propagation time of arrival from the b-th sound source to the i-th device can be derived by $\frac{\|\mathbf{r}_b - \mathbf{m}_i\|}{c}$,

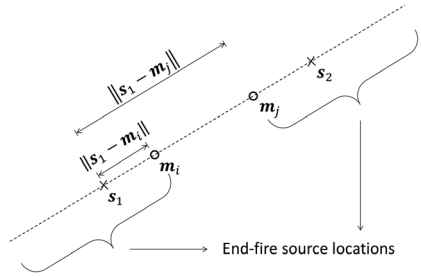


Fig. 2. Illustration of the end-fire source locations for two device locations $\mathbf{m}_i$ and $\mathbf{m}_j$ and two sound source locations $\mathbf{s}_1$ and $\mathbf{s}_2$. If a straight line is drawn to intersect both devices, the end-fire source locations are all points that lie on that line except the points that lie between the two devices.

where c is the speed of sound. Since devices and sources are distributed in various locations, the physical propagation times from sound sources to devices are different. The recordings from each device are asynchronous with unknown start times $T_i, i = 0, \cdots, N-1$. The pairwise time-shift of two devices is denoted as $T_{ij} = T_i - T_j$. Considering both propagation times and unknown start times, the time difference of arrival of the b-th source between the i-th and j-th device can be expressed as

$$\tau_{ijb} = \frac{\|\mathbf{r}_b - \mathbf{m}_i\| - \|\mathbf{r}_b - \mathbf{m}_j\|}{c} + T_{ij}. \tag{2}$$

According to the reverse triangle inequality $\|\mathbf{r}_b - \mathbf{m}_i\| - \|\mathbf{r}_b - \mathbf{m}_j\| \leq \|\mathbf{m}_i - \mathbf{m}_j\|$, the absolute TDOA values are upper-bounded by the inter-device distance. The extreme TDOAs are achieved when sources reside at end-fire locations (Fig. 2).

Suppose we have two sources on the left side and right side of the end-fire locations, respectively. The extreme TDOAs can be expressed as [10]

$$\tau_{ij}^{min} = -\frac{\|\mathbf{m}_i - \mathbf{m}_j\|}{c} + T_{ij}, \tag{3}$$

and

$$\tau_{ij}^{max} = \frac{\|\mathbf{m}_i - \mathbf{m}_j\|}{c} + T_{ij}. \tag{4}$$

Given the known speed of sound c, the inter-device distance d_{ij} can be calculated from the maximum and minimum TDOAs as

$$d_{ij} = \frac{c}{2}(\tau_{ij}^{max} - \tau_{ij}^{min}). \tag{5}$$

In this way, the distance between two devices can be estimated using TDOA information, even when their relative time-shift is unknown.

For N devices there are $\frac{N(N-1)}{2}$ device pairs. Given the distance of each device pair calculated using (5), the relative device positions $\mathbf{M}$ can be calculated by the closed-form position estimator expressed by [45], [46]

$$\mathbf{M} = \mathbf{U}_P \cdot \mathbf{X}_P{}^{\frac{1}{2}} \cdot \mathbf{Q}, \tag{6}$$

where $\mathbf{Q}$ is a $P \times P$-dimensional orthogonal rotation matrix. $\mathbf{U_P}$ and $\mathbf{X_P}$ are calculated by the best rank-P approximation of the singular value decomposition of the symmetric matrix

TABLE II
IMPORTANT NOTATIONS USED IN THIS PAPER

Symbol	Definition
a_i,A_i	Time- and STFT representation of i-th channel recording
a_i^w,A_i^w	Time- and STFT representation of wth segment a_i
B_i, B_i^w	Local time-frequency peaks of A_i and A_i^w
c	Sound speed
d_{ij}, $\tilde{d}_{ij}$	Ground-truth and estimated distance between i-th and j-th device
F	Hashed landmark value
s_b, S_b	Time- and STFT representation of b-th sound source
g_{ib}	Attenuation from b-th source to i-th device
t_{ib}, n_{ib}	Propagation time and down-sampled delay
n_G, λ_G, λ_E	Threshold parameters of baseline method
G	Gaussian function for pruning
L	Audio segment length
$\tilde{N}_{ij}^w$	Number of matched landmarks in wth segment
$\tilde{N}_{ij}$	Average number of matched landmarks across all segments
r_{ij}	Generalised cross correlation between i-th and j-th channels
R	STFT hop size
T_{ij}	Time offset between i-th and j-th device
q	qth-quantile
α	Threshold for outlier TDOA removal
γ	Pruning thresholding surface
β,σ	Parameters for controlling pruning threshold surface γ
δ, τ	Time shift and time delay
τ_{ijb}	TDOA of b-th source between i-th and j-th device
τ_{ij}^{min}, τ_{ij}^{max}	Min. and max. TDOA between i-th and j-th device
$\tilde{\tau}_{ij}^{min}$, $\tilde{\tau}_{ij}^{max}$	Estimated min. and max. TDOA
$\mathbb{F}_i^w(n)$	Set of hashed landmark values of A_i^w at frame n
$\mathbb{T}_{ij}$	Set of TDOAs between i-th and j-th recordings
$\mathbf{m}_i$, $\tilde{\mathbf{m}}_i$	Ground-truth and estimated location of i-th device
$\mathbf{M}$, $\tilde{\mathbf{M}}$	N Ground-truth and estimated device locations matrix
$\mathbf{Q}$	Rotation matrix for device localisation
$\mathbf{r_b}$	Location of b-th sound source
$\mathbf{R}$	Source location matrix
$\mathbf{y}$	Audio landmark formed by two local peaks

$\frac{1}{2}[\dot{\mathbf{d}} \cdot \mathbf{J}^T + \mathbf{J} \cdot \dot{\mathbf{d}}^T - \dot{\mathbf{D}}]$, where $\mathbf{J} = [1, \cdots, 1]^T_{(N-1)\times 1}$, $\dot{\mathbf{d}} = [d^2_{0,1}, \ldots, d^2_{0,N-1}]^T_{(N-1)\times 1}$, and

$$\dot{\mathbf{D}} = \begin{bmatrix} d^2_{1,1} & \cdots & d^2_{1,N-1} \\ \vdots & \ddots & \vdots \\ d^2_{N-1,1} & \cdots & d^2_{N-1,N-1} \end{bmatrix}_{(N-1)\times(N-1)}.$$

This closed-form estimator has been shown to achieve 1.5 times the Cramer-Rao Lower Bound when the interdevice distances are corrupted by additive Gaussian noise [46]. Due to the rotation matrix $\mathbf{Q}$, the obtained geometrical configuration of the array is invariant against rotation, translation and reflection. This is an inherent limitation in both the employed closed-form estimator (6) and the well-known multidimensional scaling algorithm [44].

This device localization framework typically requires end-fire sources with respect to each pair of devices to estimate the minimum and maximum TDOAs. This requirement can be satisfied in some real-world scenarios. For example, when a group of people simultaneously record a public event using mobile devices (smartphones), some environmental sounds (e.g. people chatting or cheering and cars passing by) can be regarded as end-fire sources.

The next task is to estimate the extreme TDOAs from multiple sources. Important notations used in this paper are listed in Table II.

B. Baseline Solution for Extreme TDOA Estimation

The baseline solution [10], [11] estimates extreme TDOAs with traditional GCC-PHAT methods. The algorithm is briefly summarized below, using two microphones i and j as an example.

First, STFT is applied to the audio streams, obtaining $A_i(n,k)$ and $A_j(n,k)$, where n and k are the frame and frequency indices, respectively. A speech-to-noise ratio (SNR) based voice activity detector (VAD) is applied to detect the frames with active sound. A frame is flagged as active if its SNR is over a threshold λ_E. Next, GCC-PHAT is applied in each active frame to calculate the generalized cross-correlation function between two microphones i and j:

$$r_{ij}(n,\tau) = \sum_{k=1}^{L_k} \frac{A_i(n,k)A_j^*(n,k)\mathrm{e}^{j2\pi f_k \tau}}{|A_i(n,k)A_j^*(n,k)|}, \quad (7)$$

where τ is the time delay, L_k is the total number of frequency bins in the whole frequency band, and f_k is the frequency at the k-th frequency bin. Assuming at most one source is active in the n-th frame, its TDOA is estimated as

$$\hat{\tau}_{ij}(n) = \arg\max_{\tau} r_{ij}(n,\tau), \quad (8)$$

where τ can be searched in the whole time frame.

A gating procedure is applied to remove the outliers of the TDOA estimation $\hat{\tau}_{ij}(n)$ in all frames. In this gating procedure, a TDOA value is flagged as an outlier if it differs more than λ_G samples between any n_G frames after.

After outlier removal, the remaining TDOA values are sorted to find the minimum and maximum TDOAs. A q-quantile operator is used to improve the robustness to residual outliers. Specifically, the minimum and maximum TDOAs are chosen to be the first $\lfloor (1-q)N_t \rceil$ and $\lfloor qN_t \rceil$ elements in the sorted TDOA set, respectively, where q is in the range [0,1], but close 1, N_t is the number of remaining TDOAs, and $\lfloor \cdot \rceil$ denotes the nearest integer.

IV. PROPOSED FINGERPRINTING BASED EXTREME TDOA ESTIMATION

A. Audio Landmark and Single-Source TDOA

Landmark-based audio fingerprinting is generally used for coarsely synchronizing audio recordings [12], [49], [50]. However, the extracted landmark features contain some valuable information about the TDOA information of the sound sources. Without loss of generality, we consider two microphones i and j.

The classical audio landmark fingerprinting converts a time-domain signal $a_i(t)$ into a sparse high-dimensional discrete-time landmark feature set $\mathbb{F}_i(n)$ [12]. At first, the time-domain signal $a_i(t)$ is transformed using the short-time Fourier transform (STFT) $A_i(n,k)$, where k is the frequency index and n is the frame index, which downsamples the time axis via the STFT hop size R. Next, local spectral peaks $B_i(n_p,k_q)$ are selected from the power spectral amplitude $|A_i(n,k)|^2$ by comparing it with a threshold surface $\gamma(n,k)$, where n_p and k_q denote the frame and frequency indices of the detected local peak, respectively. The threshold is initialized by the peaks found in the first few frames. Then at each frame n is updated by a decaying factor β and is also raised by peaks found in the previous

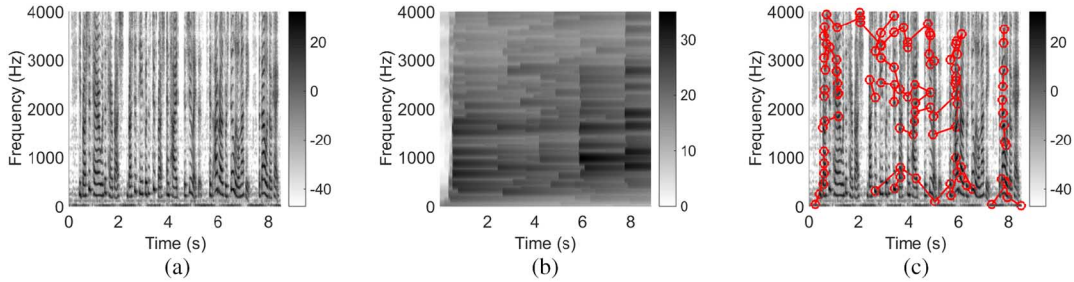


Fig. 3. Visualization of audio fingerprint extraction. (a) Spectrogram of the signal. (b) Threshold (pruning) surface γ. (c) Extracted audio landmarks.

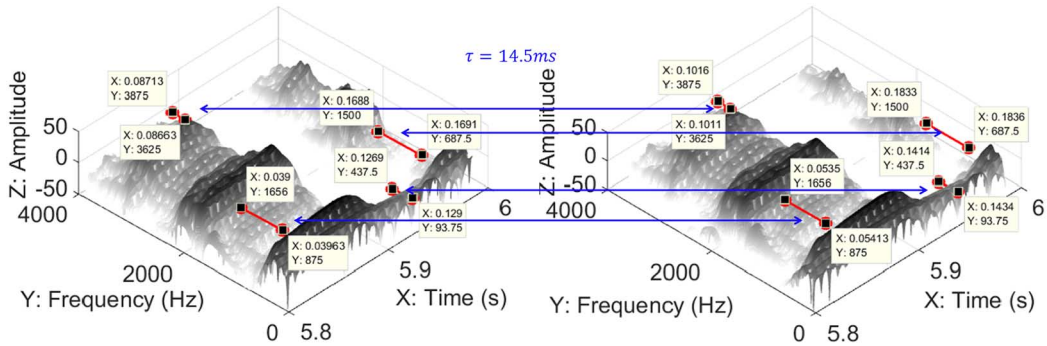


Fig. 4. Visualization of audio fingerprint matching between two audio channels with a time offset of $0.0145s$. Four examples of matched landmarks are drawn on the corresponding spectrograms. Each pair of matched landmarks shows a time delay of $\tau = 0.0145s$.

frame. All the local peak values higher than the threshold are kept in $B_i(n,k)$, with other peaks set to zeros ('pruning') [51]. The threshold is updated as

$$\gamma(n,k) = \max\left(\frac{\gamma(n-1,k)}{\beta}, G(k) \circledast B_i(n,\cdot)\right), \quad (9)$$

where $G(k)$ is a Gaussian function and $\circledast$ denotes the convolution operator in frequency. The number of local peaks is controlled by β and the variance, σ, of $G(k)$. The larger β and σ, the more local peaks will be selected.

A landmark, $\mathbf{y}(n_1,k_1;n_2,k_2)$, is formed by pairing up two nearby local spectral peaks $B_i(n_1,k_1)$ and $B_i(n_2,k_2)$. To reduce the dimension, each landmark is hashed into an integer value using $F = k_1 2^{12} + (k_2 - k_1)2^6 + (n_2 - n_1)$ [52]. In this way, the obtained landmarks associated with the time frame n are represented as a time-indexed feature set $\mathbb{F}_i(n) = \{F_u\}_{u=1}^{U_n}$, where U_n is the total number of landmarks at the frame n.

The extracted audio landmarks contain the TDOA information of the sound sources. Assume only the b-th source is active and the time offset between two microphones is zero. Then the STFT of $a_i(t)$ and $a_j(t)$ can be expressed

$$\begin{cases} A_i(n,k) = g_{ib} S_b(n - n_{ib}, k) \\ A_j(n,k) = g_{jb} S_b(n - n_{jb}, k) \end{cases}, \quad (10)$$

where $S_b(\cdot)$ is the STFT of $s_b(\cdot)$, $n_{ib} = \lfloor t_{ib}/R \rfloor$, and $n_{jb} = \lfloor t_{jb}/R \rfloor$, where the operator $\lfloor\cdot\rfloor$ denotes the integer part.

By landmark matching between the two channels [12], the landmarks corresponding to the same time-frequency peak pairs can be extracted:

$$\boldsymbol{y}_i(n_1 - n_{ib}, k_1; n_2 - n_{ib}, k_2) = \boldsymbol{y}_j(n_1 - n_{jb}, k_1; n_2 - n_{jb}, k_2), \quad (11)$$

and consequently

$$\mathbb{F}_i(n) = \mathbb{F}_j\left(n - \left\lfloor \frac{\tau_{ijb}}{R} \right\rfloor\right), \quad (12)$$

where $\mathbb{F}_i(\cdot)$ and $\mathbb{F}_j(\cdot)$ denote the extracted audio fingerprints of $a_i(\cdot)$ and $a_j(\cdot)$, respectively; and $\boldsymbol{y}_i(\cdot)$ and $\boldsymbol{y}_j(\cdot)$ denote two matched local peak pairs in the two channels. The time delay between two channels and the matched landmarks are clearly related in (12), where the hop size R determines the resolution of the time delay.

As example, in a simulated anechoic environment a sound source (speech) is placed at an end-fire location with respect to a pair of (synchronized) microphones which are $5m$ apart. The audio fingerprint extraction procedure is visualized in Fig. 3, where the spectrogram of the speech signal, the pruning threshold surface, and the extracted audio landmarks in the first channel are depicted in the three subfigures. The audio fingerprint matching results between two channels are visualized in Fig. 4, where four examples of matched audio landmarks in a short segment ($5.8s - 6.0s$) in the two channels are depicted. The matched landmarks in the two channels typically occur at the same frequency bins but with a temporal offset of $0.0145s$, which equals to the acoustic transmission time of $5m$.

B. Proposed Extreme TDOA Estimation Method

Based on the analysis above, we propose an audio-fingerprinting based method to estimate the extreme TDOAs from a multi-source environment and then utilize the estimated pairwise distances to compute the device locations (Fig. 5).

The signals $a_i(t)$, $i = 0, \cdots, N-1$ are divided into W non-overlapping segments, $a_i^w(t), w = 1, \cdots, W$, of length L. The time domain signal $a_i^w(t)$ is transformed to the audio landmark

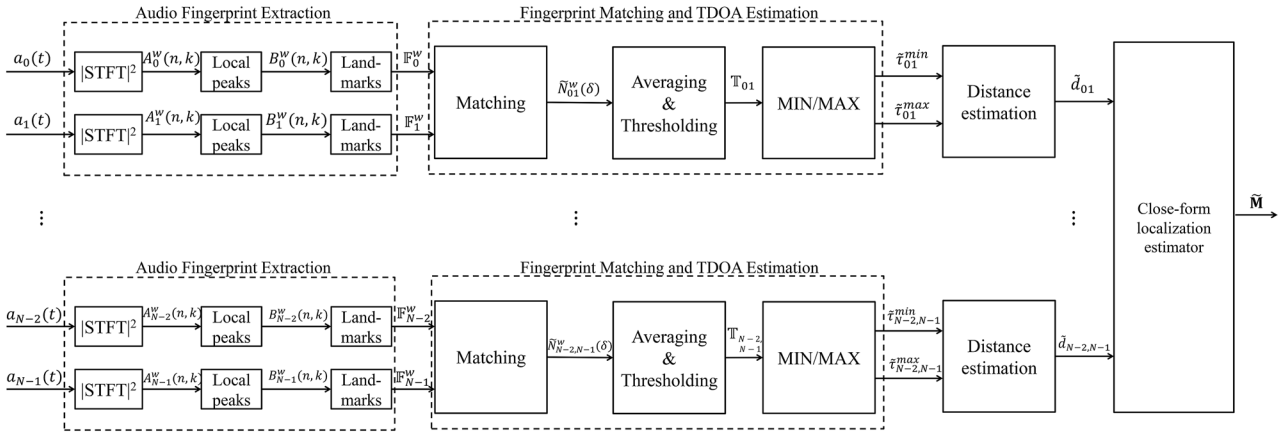


Fig. 5. The block diagram of the proposed audio-fingerprinting-based device localization method.

feature set $\mathbb{F}_i^w(n)$. In each segment we calculate the number of matched audio landmarks of $\mathbb{F}_i^w(n)$ and $\mathbb{F}_j^w(n)$ at different time shifts δ:

$$\tilde{N}_{ij}^w(\delta) = \sum_{n=1}^{\lfloor L/R \rfloor - 1} \left\langle \mathbb{F}_i^w(n) \cap \mathbb{F}_j^w(n+\delta) \right\rangle, \tag{13}$$

where $\cap$ is the intersection and $\langle\cdot\rangle$ is the cardinality of a set [12]. The number of matched landmarks $\tilde{N}_{ij}^w(\delta)$ is averaged across all the segments, obtaining

$$\tilde{N}_{ij}(\delta) = \frac{1}{W} \sum_{w=1}^{W} \tilde{N}_{ij}^w(\delta). \tag{14}$$

A small number of outliers randomly distributed across different segments may still exist in $\tilde{N}_{ij}^w(\delta)$ due to some mismatched landmarks. Averaging $\tilde{N}_{ij}^w(\delta)$ over all segments helps to suppress these outliers. We refer to $\tilde{N}_{ij}(\delta)$ as the matching score between two channels.

When only one source exists, e.g. the b-th source, its TDOA can be estimated in a similar way as GCC, i.e.,

$$\tau_{ij} = \arg\max_{\delta} \left\{ \tilde{N}_{ij}(\delta) \right\} \cdot R, \tag{15}$$

where δ is in the range $\{-\lfloor \frac{L}{R} \rfloor, \cdots, \lfloor \frac{L}{R} \rfloor\}$ with a step of 1. When multiple sources exist, $\tilde{N}_{ij}(\delta)$ is seen as a histogram of the TDOAs of these sources. Similarly to [11], we remove the residual outliers by applying a threshold α to $\tilde{N}_{ij}(\delta)$ and estimate a set of TDOAs $\mathbb{T}_{ij}$ by using

$$\mathbb{T}_{ij} = \left\{ R \cdot \delta | \tilde{N}_{ij}(\delta) > \alpha \right\}. \tag{16}$$

The minimum TDOA $\tilde{\tau}_{ij}^{min}$ and maximum TDOA $\tilde{\tau}_{ij}^{max}$ can be estimated as

$$\tilde{\tau}_{ij}^{min} = \min(\mathbb{T}_{ij}) \text{ and } \tilde{\tau}_{ij}^{max} = \max(\mathbb{T}_{ij}). \tag{17}$$

The inter-device distance $\tilde{d}_{ij}$ is calculated using $\tilde{\tau}_{ij}^{min}$ and $\tilde{\tau}_{ij}^{max}$, as given in (5). In the same way the distance of each pair of devices in the ad-hoc array can be estimated. The pair-wise distances are further used to recover the geometrical configuration of the array, based on the closed-form estimator (6).

C. Discussion

The proposed method and the baseline method calculate the correlation coefficient or the matching score between two channels in order to estimate the time delay. The baseline method exploits the phase information of the STFT signals, while the proposed method exploits the amplitude information of the STFT signals, which is more robust to environmental noise. The performance of the proposed method is mainly influenced by two classes of factors: algorithmic and acoustic factors.

1) Algorithmic Factors: The STFT hop size R plays an important role in the precision of the audio-fingerprinting-based distance estimation algorithm. As indicated in (12), the resolution of the audio-fingerprinting-based TDOA estimation is confined by R. A hop size as small as $R = 2$ (equals 250 μs@8 kHz) is used in the proposed algorithm so that an improved temporal analysis resolution is achieved and the TDOAs of different sources can be distinguished from each other in the histogram (14) as different peaks. This is in contrast to the choice in traditional audio fingerprinting techniques which have been applied to video synchronization [49] [50] or music information retrieval [53]. In these applications, the hop size R is usually chosen to be a value within the range $256 - 512$ (equals $0.032 - 0.064$ s@8 kHz), which is already enough for coarsely synchronizing audio channels but far below the requirement for TDOA and distance estimation. By employing fine-resolution audio fingerprinting, the proposed method is able to extract the TDOA information that is embedded in the audio landmarks. This is an important contribution of the proposed method.

2) Acoustic Factors: The performance of the extreme TDOAs estimation is affected by four factors: inter-device distance, interfering sources, deviation of end-fire source locations, and environment reverberation. For a sound source located at the end-fire direction of two devices, the inter-channel intensity ratio varies with the *inter-device distance*. When this distance is increased, the sound pressure at the far-end device will decrease relative to the close-end device, making it difficult to find enough matched landmarks among two channels. This influence stands out especially when the sound sources are in the near field, i.e., the source-device distance and the

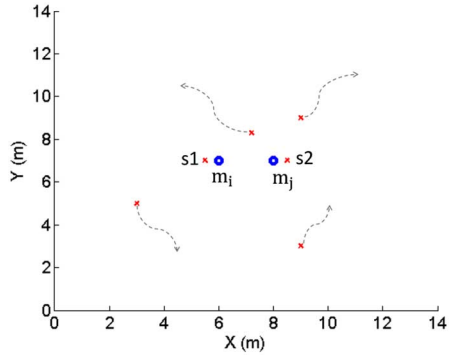


Fig. 6. Illustration of the simulation environment: a microphone pair ($\mathbf{m}_i$, $\mathbf{m}_j$), two end-fire sources ($\mathbf{s}_1$, $\mathbf{s}_2$), and several randomly placed interferences. The distance of the microphone pair varies depending on specific experiment while the distance between the end-fire source to its close-end microphone is always $0.5m$.

inter-device distance are comparable. The spectrum of *interfering sources* (i.e. sources are not in the end-fire locations) will disturb detection of spectrum peaks of the end-fire sources, decrease the number of matched landmarks, and reduce the amplitudes of desired peaks in a TDOA histogram. Ideally, the end-fire sources should be located on the line connecting two devices. In practice, the *locations of end-fire sources* may deviate from the desired ones. As a result, the estimated extreme TDOAs will also deviate from the desired values, leading to inaccurate inter-device distance estimates. The proposed method is derived with a free-field model. However, in practical applications, the influence from the *environment reverberation* cannot be neglected. The reverberation typically generates spurious images of the sound sources [54], [55], degrading the extreme TDOA estimation performance. The specific influence of the above factors will be investigated in Section V-B.

D. Parameter Selection

We use the shoebox simulator [56] to generate different acoustic scenarios for parameter selection and performance evaluation. We set the sampling rate to $8kHz$ and sound speed to $345m/s$. The simulated enclosure is of size $14m \times 14m \times 4m$, as shown in Fig. 6, with the reverberation time controlled by varying the absorption coefficients of the walls. We use reverberation time 0 except for the Acoustic Scenario 5 in Section V-A. A pair of microphones together with a pair of end-fire sources are placed in the center of the room. The distance of the two microphones varies depending on the specific acoustic scenario that is used. The two end-fire sources are always placed $0.5m$ away from the two microphones. Several interfering sources are placed randomly around the microphones. The number of interfering sources also depends on specific acoustic scenarios. Usually, the end-fire sources are chosen from male or female speeches while the interfering sources are chosen from speech, traffic, bird, or white noise sounds. All the sound files (end-fire sources and interfering sources) are of similar intensities.

The proposed distance estimation method has five parameters: the audio processing segment length L, the hop size R of the STFT analysis, the decay rate β and the variance σ of the threshold surface in (9), and the outlier threshold α in (16).

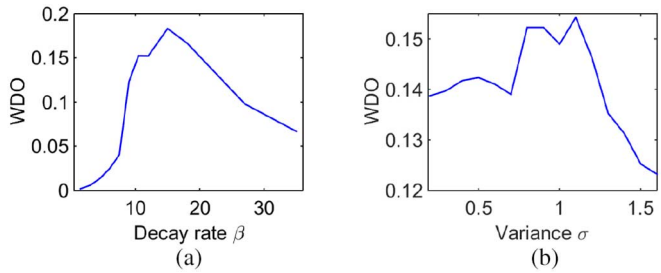


Fig. 7. The W-disjoint orthogonality (WDO) measure for different (a) decay rates β and (b) variances σ.

The length of the processing segment L controls the TDOA searching range. We set it as $0.2s$, which is equivalent to a propagation time between two devices of about $70m$ apart, i.e. the maximum allowed inter-device distance in the algorithm is $70m$. The hop size R controls the STFT temporal analysis resolution and the TDOA estimation precision. As discussed in Section IV-C, we set $R = 2$ samples, which represents a temporal resolution of $0.00025s$. The decay rate β and the variance σ in (9) control the amount of detected local spectral peaks [12]. Using a larger value of β and σ may increase the number of detected local peaks and landmarks, whereas the number of falsely matched landmarks will also rise. Thus a trade-off between the quantity and quality of the matched landmarks has to be made when determining the values of β and σ. We employ the W-disjoint orthogonality (WDO) [13] to measure the ratio between energies of the desired signals (end-fire sources) and the interferences at the time-frequency bins where the matched landmarks are located. The WDO measure is defined as

$$\text{WDO} = \frac{1}{L_n L_k} \sum_{n=1}^{L_n} \sum_{k=1}^{L_k} \frac{\|M(n,k)A_S(n,k)\|^2 - \|M(n,k)A_I(n,k)\|^2}{\|A_S(n,k)\|^2}, \tag{18}$$

where L_n and L_k are the total number of time frames and frequency bins, respectively; $M(n,k)$ at the (n,k)-th bin is a binary time-frequency mask which is set to 1 when the bin contains the matched landmark and set to 0 otherwise; $A_S(n,k)$ and $A_I(n,k)$ are the STFTs of the desired signals and the interferences at the (n,k)-th bin, respectively. We carry out a simulated experiment to investigate how WDO varies with β and σ using one interference (car sound). All the sound files are $12s$ long. The distance between the two microphones is $4m$. The WDO measures are calculated at the two microphones, respectively, and then averaged. The results are shown in Fig. 7(a) and Fig. 7(b) for β and σ, respectively. In Fig. 7(a), σ is fixed at 1, and β is varied from 1.5 to 35. In Fig. 7(b), β is fixed at 30, and σ is varied from 0.2 to 1.6. $\sigma = 1$ and $\beta = 30$ are the default values suggested by [52]. For both β and σ, the WDO measure at first increases with the increasing parameter value when more landmarks are detected. However, after reaching a peak value the WDO measure starts to drop with the increasing parameter value because more mismatched landmarks are found. We therefore choose $\beta = 15$ and $\sigma = 1.1$, which maximize the WDO measures.

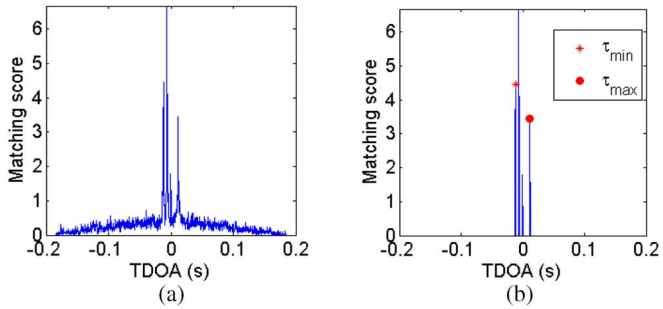


Fig. 8. An example of estimating extreme TDOAs from the TDOA histogram. The microphone distance is $4m$. (a) Matching score ($\tilde{N}_{ij}(\tau)$) before thresholding. (b) Matching score after thresholding.

The threshold α in (16) removes the outliers of the mismatched landmarks. Usually, after the averaging processing (14) across audio segments the outliers have already been effectively suppressed. Thus, the proposed method is not sensitive to α and we choose it to be between 1.5 and 2.

The above values of α, β and σ are selected for anechoic scenarios. In reverberant scenarios, which are not the main focus of this paper, the optimal values of these parameters can be determined in a similar way using the WDO measure.

Fig. 8 illustrates the histogram of the TDOAs obtained using the above parameter values, i.e. $L = 0.2s$, $R = 2$, $\beta = 15$, $\sigma = 1.1$, and $\alpha = 1.5$. We use two end-fire sources (a male and a female speeches) and two interference (a music and a car sounds). All the sound files are $12s$ long. Fig. 8(a) shows the matching score (cf. (14)), i.e. the average number of matched landmarks across all the segments, at different time shifts. Strong peaks can be clearly observed in the area between $-0.05s$ and $0.05s$. To extract the extreme TDOAs thresholding α is applied, with the results shown in Fig. 8(b). The peaks of the two extreme TDOAs can be observed at the time $-0.012s$ and $0.01125s$, which denote $\tilde{\tau}_{ij}^{min}$ and $\tilde{\tau}_{ij}^{max}$, respectively. Finally, the microphone distance is estimated as $\tilde{d}_{ij} = 4.01m$ using (5).

V. Experimental Results

A. Experimental Setup

Two methods are considered and compared in the experiment: the GCC-based baseline method (cf. Section III-B) and the proposed audio-fingerprinting-based (AF-based) method. The audio fingerprint/landmark extraction algorithm is implemented using the code in [52]. The specific parameters of the two methods are summarized in Table III. The parameters used in the baseline method are set based on [10], while the parameters used in the proposed method are set based on the discussion in Section IV-D These parameter values are used throughout the experiment unless otherwise stated.

The relative error, e_{RE}, and the root-mean-square error, e_{RMS}, are used to evaluate the inter-device distance estimation performance and the device localization performance, respectively. Given the true inter-device distance d_{ij} and the estimated value $\tilde{d}_{ij}$, e_{RE} is defined as

$$e_{RE} = \frac{|\tilde{d}_{ij} - d_{ij}|}{d_{ij}} \times 100\%. \tag{19}$$

TABLE III
Parameters used in the Baseline and Proposed Methods

Method	Name	Symbol	Value
Baseline	Window type, length, hop size	-	Hamming, 2048, 1024 (samples@$8kHz$)
	VAD threshold	λ_E	0.1
	Gating threshold I	λ_G	6
	Gating threshold II	n_G	2
	q-quantile	q	0.95
Proposed	Window type, length, hop size	-,-,R	Hamming, 128, 2 (samples@$8kHz$)
	Segment length	L	$0.2s$
	Decay rate	β	15
	Variance	σ	1.1
	Outlier threshold	α	2

We assume that the estimation failed when the relative error is larger than 100% and thus set this value as an upper bound of the calculated relative error. The device locations are estimated from the pair-wise distances of the devices using the closed-form estimator (6). As mentioned in Section III-A, the geometrical configuration of the array obtained by the closed-form estimator is not invariant against rotation. To calculate the error between the estimated device locations and the ground-truth locations, we compute the rotation matrix $\mathbf{Q}$ using the ground-truth device distances and locations and apply it to the estimated device locations [45], [46]. Given the true location of N devices $\mathbf{m}_i$, $i = 1, \cdots, N$, the estimated value $\tilde{\mathbf{m}}_i$ and the ground-truth rotation matrix $\mathbf{Q}$, the root-mean-square (RMS) error is used to evaluate the device location estimation performance:

$$e_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{Q}\tilde{\mathbf{m}}_i^T - \mathbf{m}_i^T\|^2}. \tag{20}$$

Two other measures are used: signal-to-interference ratio (SIR) and direct-reverberation-ratio (DRR). SIR, which measures the noise density of the acoustic environment, is

$$\text{SIR} = 10\log\frac{P_S}{P_I}, \tag{21}$$

where P_S denotes the sum of the powers of the two end-fire sources while P_I denotes the sum of the powers of the interfering sources. DRR, which measures the reverberant density of the acoustic environment, is

$$\text{DRR} = 10\log\frac{P_D}{P_R}, \tag{22}$$

where P_D denotes the sum of the powers of the direct sounds from the two end-fire sources while P_R denotes the sum of the power of the reverberant sounds from the two end-fire sources.

The following five simulation scenarios are designed to evaluate the performance of the two algorithms for various device distances, SIR levels, interfering source number, end-fire source location deviation and reverberation time. The simulator described in Section IV-D is used in the simulation. In all scenarios, we always have two end-fire sources (male and female speech) close to the two microphones. In the first four scenarios, we only consider anechoic environments.

Scenario 1-Different Inter-Device Distances: Two end-fire sources and one interference (traffic noise) are used and placed as shown in Fig. 6. Nine device distances from $1m$ to $9m$ with an interval of $1m$ are tested. The interference is placed randomly around the two microphones. For each device distance, 20 realizations of random interference positions are used. The length of the sound files is $10s$.

Scenario 2-Interference with Different Intensities: The device distance is $2m$. Two end-fire sources and one interference (Gaussian white noise) are used. The intensity of the interference is varied so that the average signal-to-interference ratio (SIR) at the two microphones varied from $-10dB$ to $25dB$ with an interval of $5dB$. For each SIR, 20 realizations of random positions of the interference are used. The length of the sound files is $10s$.

Scenario 3-Different Number of Interferences: The device distance is $2m$. Two end-fire sources and different number of interferences, varying from 2 to 10, are used. The sound files for the interferences are randomly selected from human, music, bird and traffic noise. The length of the sound files is $10s$. For each interference number, 20 realizations of random interference placement are used. With all the sound files (end-fire sources and interferences) having the same intensity, the average SIR at the two microphones varies from -5 to $3dB$, depending on the number of interferences.

Scenario 4-Deviation of End-fire Source Location: The device distance is $2m$. Two end-fire sources and no interference are used. The placement of the end-fire sources deviates from the desired locations. The deviation is set as $\pm\Delta$ along the x-, y-, and z-coordinates, respectively. The value of Δ varies from $0.2m$ to $0.6m$ with an interval of $0.1m$. For each Δ, all the combinations of the deviations along the three coordinates are used.

Scenario 5-Different Reverberation Times: The configuration in Scenario 3 is used, with two different numbers of interferences (0 and 4). Since the size of the simulated room is very big in the simulator (e.g. $14m \times 14m$), we use different reverberation times varying from $0.4s$ to $3s$. For reference, the averaged direct-to-reverberation ratios (DRRs) of the two end-fire sources are also calculated.

B. Performance Comparison in Simulated Environments

1) Testing Signal Length: The length of the signal used for TDOA estimation also influences the performance of both the baseline and the proposed method. Specifically, the end-fire sources should be active long enough so that they could be reliably detected from the TDOA histogram. To investigate the influence of the signal length, we use the configuration in the simulated Acoustic Scenario 3 with different numbers of interferences: 2, 5, and 10. One realization is tested. All the sound files are set to be of the same length, ranging from $2s$ to $64s$.

The distance estimation performance of the baseline and the proposed method using different data lengths is shown in Fig. 9. Both methods perform worst when a short signal length of $2s$ is used. Their performance improves significantly when the signal length is increased from $2s$ to $8s$ and saturates afterwards. The results in Fig. 9 are obtained for various scenarios (i.e. using different number of interferences (2, 5, and 10)), and the conclusion is consistent for all these scenarios. Based on this analysis, we choose a data length of $10s$ in other experiments, as described in Section V-A.

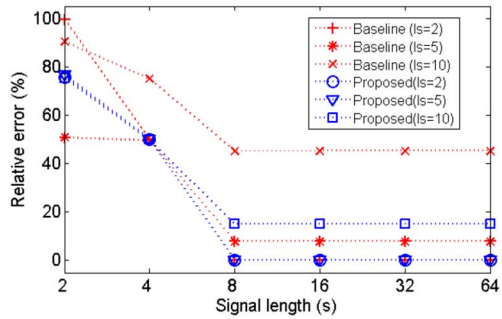


Fig. 9. Distance estimation performance by the baseline and the proposed methods using different signal lengths. Inter-device distance $2m$, different number of interferences (Is $=$ 2, 5, 10) are tested.

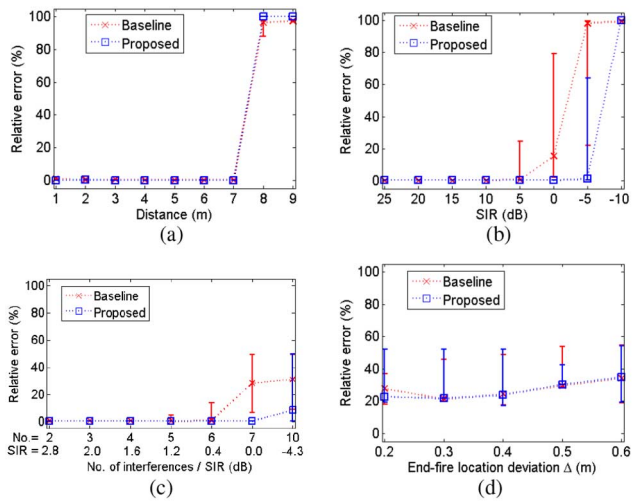


Fig. 10. Distance estimation performance of the baseline and the proposed methods: (a) Scenario 1: different inter-device distances, 1 interferences; (b) Scenario 2: different intensities of interference white Gaussian noise, inter-device distance $2m$, 2 interferences; (c) Scenario 3: different number of interferences, inter-device distance $2m$; (d) Scenario 4: different deviations of end-fire source locations, inter-device distance $2m$, 0 interferences.

2) Simulation Results in Anechoic Environments: The performance comparison of the baseline and the proposed methods in different scenarios in anechoic environments is given in Fig. 10. In each panel, the distance estimation results are plotted using the median value of the 20 realizations while the error bar shows the first and third quartiles.

The experimental results of Scenario 1 in Fig. 10(a) show that with only one interference both the baseline and proposed methods can accurately estimate the inter-device distance up to $7m$. However, both methods fail when the inter-device distance is equal to or larger than $8m$. With the increase of device distance, the inter-channel intensity of the end-fire sources becomes smaller while the influence of the noise becomes dominant. As a result, fewer landmarks that belong to the end-fire sources can be detected and matched between the two channels. In the obtained TDOA histogram, the peaks of the end-fire sources become obscured by the peaks of the noise signals, making it difficult to detect the correct extreme TDOA values.

The experimental results of Scenario 2 in Fig. 10(b) show that the baseline and the proposed methods can accurately estimate the inter-device distance at high SIRs ($\geq 10dB$). However, both methods show degraded performance when the SIR is decreased from $5dB$ to $-5dB$, and both fail when the SIR equals $-10dB$. In the SIR range from $-5dB$ to $5dB$, the proposed method outperforms the baseline method. Although they achieve similar performance in terms of the median values of the estimation errors (in some cases the proposed method achieves even lower median errors), the proposed method has much fewer outlier estimates than the baseline method. The superior performance of the proposed method becomes more evident with the increase of the noise level.

The experimental results of Scenario 3 in Fig. 10(c) show that the baseline and the proposed methods can accurately estimate the inter-device distance when the number of interferes is smaller than 5. However, both methods show degraded performance when the number of interferences is increased from 5 to 10. When the number of interferences is between 5 and 10, the proposed method clearly outperforms the baseline method with lower median errors and less outlier estimates. For reference, the SIRs at the microphones are also indicated in Fig. 10(c). The results are consistent with those observed in Fig. 10(b): the benefits of the proposed method are clearly observed in the SIR range from 0 to -5 dB.

The experimental results of Scenario 4 in Fig. 10(d) show that the performance of both methods degrades when the location deviation of the end-fire sources is increased. Since both methods use the estimated extreme TDOAs to calculate the device distance, the deviation of the end-fire source locations imposes similar influence on them (for both methods the median value of the estimation errors is about 20% when the location deviation rises up to $\pm 0.4m$).

In summary, the proposed method performs similarly to the baseline method in an acoustic environment with low noise, but outperforms it in an environment with heavy noise, typically with an SIR lower than $0dB$. The GCC-based baseline method operates in a frame-wise manner in the time domain with the assumption that there are always some time frames where the end-fire sources are dominant. At each frame it only estimates one dominant TDOA and the extreme TDOAs are detected from the histograms of the dominant TDOAs in all the time frames. If the end-fire source is always weaker than the interfering sources, the baseline method will fail since it can not detect the frames with extreme TDOAs. In contrast, the proposed audio-fingerprinting-based algorithm is not constrained by this assumption. By locating the matched landmarks formed by local time-frequency peaks, it can robustly detect the extreme TDOAs even when the end-fire sources are always weaker than the interfering sources. As shown in the simulation results, when the SIRs in the microphones are between $-5dB$ and $0dB$, the proposed method performs much more robustly than the baseline method, with much fewer outlier estimates.

Simulation Results in Reverberant Environments: The performance of the baseline and proposed methods is compared in different reverberant densities with reverberation time varying from $0.4s$ and $3s$ (Fig. 11). For reverberation time increasing from $0.4s$ to $3s$, the DRR decreases accordingly from $9dB$ to $0.5dB$. Two scenarios are tested with different number of interferences. In the first scenario with no interferences, both methods can estimate the device distance accurately in all reverberant densities. The second scenario is more challenging with four interferences. While both methods can estimate the device distance accurately at high DRRs ($\geq 7dB$), their performance degrades almost linearly when the reverberation density is increased. Both methods fail when DRR equals $0.5dB$. The performance difference between the two methods can be observed in the DRR range from $7dB$ to $2dB$, where the proposed method outperforms the baseline method in most cases with lower median errors and less outliers. Although only two cases (with 0 and 4 interferences) are investigated in this experiment, the obtained results can still demonstrate that the reverberation influences the performance of both methods especially when multiple sources are active. Both methods perform well in low reverberation (e.g. DRR $> 7dB$) and fail in high reverberation (e.g. DRR $< 2dB$), but the proposed method still works more robustly than the baseline method in scenarios with medium reverberation densities. Thus, we conclude that the proposed method (which was derived with a free-field model) is potentially applicable to reverberant scenarios.

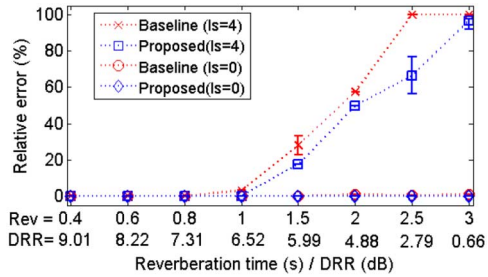


Fig. 11. Distance estimation performance of the baseline and the proposed methods in different reverberant scenarios. The inter-device distance is $2m$. Different number of the interferences are tested (Is $= 0, 4$).

C. Performance Comparison in Real Environments

We first investigate the inter-device estimation performance and then use the estimated pair-wise distances for device localization in a real environment.

The real recordings were made in a quiet public square (approximately $20m \times 20m$), with low reverberation except for reflections from the nearby buildings and the ground. The reflections captured by the microphones would increase the challenge of TDOA estimation. The temperature was about $20°C$. Four Samsung Galaxy III smartphones were placed at 4 fixed positions and used as recording devices, while the testing sound was played by a monitor loudspeaker (Genelec 8010) at 17 fixed positions and recorded individually. The real recording environment and the geometrical configuration of the devices and loudspeakers were shown in Fig. 12(a) and (b), respectively. The ground-truth inter-device distances measured by a laser distance meter (Leica Disto A2) were: $d_{01} = 5.31m$, $d_{02} = 5.60m$, $d_{03} = 3.55m$, $d_{12} = 1.47m$, $d_{13} = 6.26m$ and $d_{23} = 5.73m$.

The testing sound was composed of 18 recordings including speech, car, bird, and music files, each 10 seconds long. Mono-channel recording was used with the sampling rate of $8kHz$. The sampling rate of each device was measured in advance and

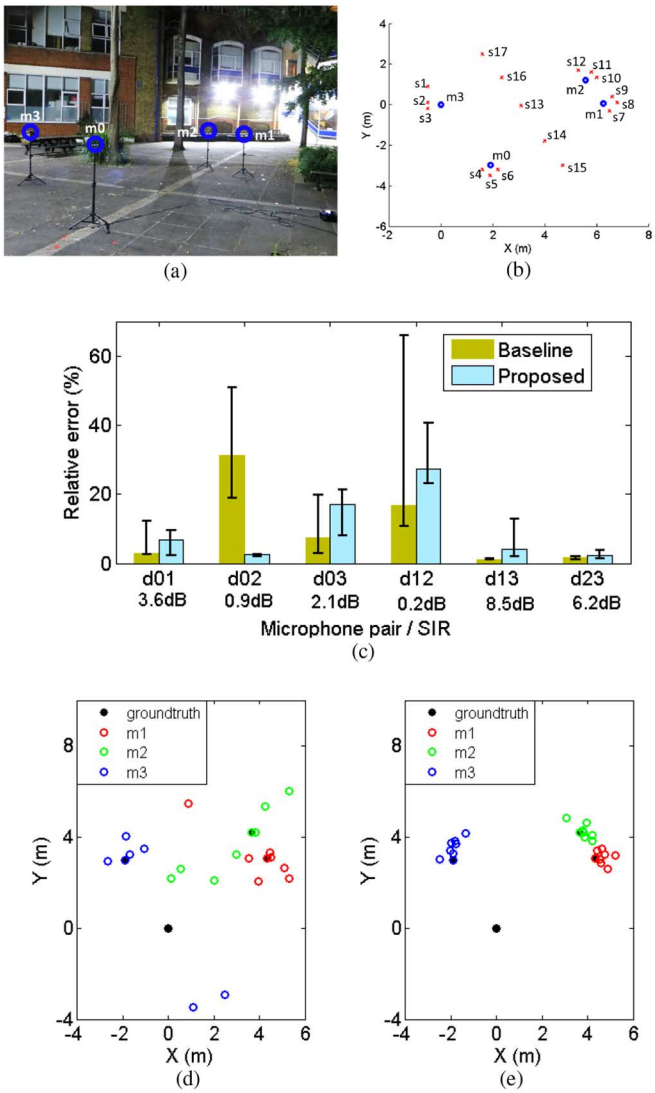


Fig. 12. Illustration of the environment and geometrical configuration, inter-device distance estimation, and device location estimation results for the real recordings. (a) The environment (a public square). (b) Geometrical configuration. (c) Inter-device distance estimation results of 6 pairs of devices. (d) Device localization result by the baseline method. (e) Device localization result by the proposed method.

resampling was applied to the recorded signals to compensate for the clock drift. To generate a multi-source environment, the recordings at individual locations were added, which contain different environment noise caused by wind and passing cars, hence the superimposition led to a noisier output than the true environment. We implement 8 realizations. In each realization, 4 end-fire sources are randomly selected from $s1 - s12$, while 2 interfering sources are randomly selected from $s1 - s17$. In this case, the end-fire sources might deviate from the desired positions. The end-fire sources use speech files while the interfering sources use files randomly chosen from speech, music, bird and car sound files.

In addition to the challenges of large inter-device distance and multiple sources, the real-recording scenario also suffers from environment noise, acoustic reflections, and non-uniform microphone sensitivities. The inter-device distance estimation results for the 6 pairs of devices are given in Fig. 12(c). The distance estimation results are plotted using the median value of the 8 realizations, while the error bar shows the first and third quartiles. For reference, the SIRs of the device pairs are also indicated in Fig. 12(c). The inter-device distance estimation performance for real recordings is consistent to those for simulations (cf. Fig. 10(c)). Specifically, the baseline method and the proposed method perform similarly at high SIRs (e.g. for the microphone pairs d_{01}, d_{03}, d_{13}, d_{23} with SIRs $3.6dB$, $2.1dB$, $8.5dB$ and $6.2dB$, respectively). At low SIRs (e.g. for the microphone pairs d_{02} and d_{12} with SIRs $0.9dB$ and $0.2dB$, respectively) the median values of the errors by the two methods are similar, but the proposed method yields much fewer outlier estimates than the baseline method, as indicated by the error bars.

The device locations are estimated from the obtained pair-wise distances of the devices using the closed-form estimator (6). To solve the rotation ambiguity problem as indicated in (20), we compute a rotation matrix $\mathbf{Q}$ from the ground truth locations and apply it to the estimated device locations. Thus, the device localization of the baseline and the proposed methods can be compared. The device localization results of the 8 realizations by the baseline and the proposed methods are given in Fig. 12(d) and (e), respectively. It can be observed that the device locations estimated by the proposed method deviates from the true locations less than those obtained by the baseline method. Finally, the average RMS errors (cf. (20)) of the proposed method and the baseline method are $0.5m$ and $2.12m$, respectively. With better inter-device estimation results, the proposed method outperforms the baseline method in terms of device localization.

In summary, while a free-field model was used during its derivation, the proposed method is potentially applicable to reverberant scenarios. In noisy anechoic simulation and open-space environments, the proposed method achieves similar inter-distance estimation accuracy as the baseline method in low-noise scenarios with an SIR higher than $5dB$, but works more robustly in noisy scenarios with an SIR ranging from $-5dB$ to $0dB$. Additional experiments in simulated reverberant scenarios show that reverberation will significantly influence the performance of both methods, especially when multiple sources are active. Both methods perform well in low reverberation (e.g. DRR $> 7dB$) and fail in high reverberation (e.g. DRR $< 2dB$). However, the proposed method still works more robustly than the baseline method in scenarios with medium reverberation densities.

D. Computational Complexity Analysis

The proposed inter-device distance estimation method consists of two main blocks, audio fingerprint extraction and matching. The first block dominates the computational complexity. The computational cost of audio fingerprint extraction, which consists of STFT analysis and landmark detection, is closely related to the hop size R of the STFT analysis. The computational cost of the STFT analysis is inversely proportional to the hop size, whereas the computational cost of landmark detection is proportional to the size of the time-frequency spectrogram. Given signal length L_s, STFT window length L_w and hop size R, the cost of the audio fingerprinting block, C_{AF}, can be expressed as

$$C_{AF} \approx \frac{L_s}{R} C_{FFT}(L_w) + \frac{L_s}{R} C_{conv}(L_w), \qquad (23)$$

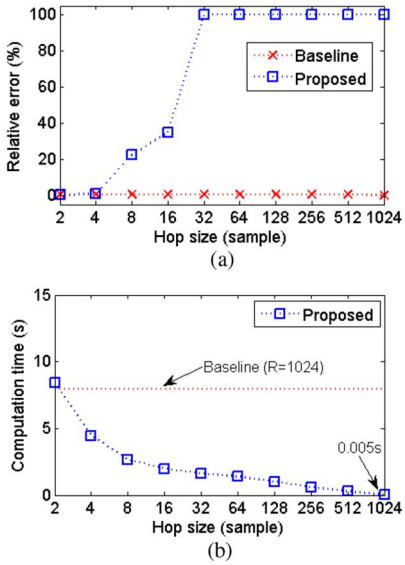


Fig. 13. (a) Distance estimation error and (b) computation time at different hop sizes by the baseline and the proposed methods. The inter-device distance is $4m$. Two end-fire sources and two interferences. The signal length is $10s$.

where $C_{FFT}(L_w)$ denotes the cost of L_w-point fast Fourier transform (FFT) analysis, and $C_{conv}(L_w)$, which dominates the computation of landmark detection, denotes the computational cost of the convolution operation per time frame in (9) and also relies on L_w.

The computation of the baseline method is dominated by STFT analysis and generalized cross-correlation calculation. The cross-correlation calculation in one frame is related to the window length of the STFT analysis and it has to be performed in each analysis frame. The cost of the GCC-based method, C_{GCC}, can be expressed as

$$C_{GCC} \approx \frac{L_s}{R} C_{FFT}(L_w) + \frac{L_s L_w^2}{R} C_{corr}, \qquad (24)$$

where C_{corr} denotes the cost for computing the fraction in (7) per time frame n, frequency k, and time-shift τ. The analysis shows how the complexity of both the baseline method and the proposed method increases when decreasing the hop size. As the computational costs C_{FFT}, C_{conv} and C_{corr}, depend on the specific implementation, we compare the computational cost experimentally.

To analyze how the distance estimation performance and the computational complexity of the proposed method varies with the hop size, we use the simulator in Section IV-D with two end-fire sources and two interferences (car sound and music). The signal length is $10s$. The microphone distance is $4m$. The baseline and the proposed method are applied using different hop sizes from 2 to 1024. The distance estimation results are given in Fig. 13(a). The performance of the baseline algorithm does not depend on the hop size, because the GCC method calculates the TDOA by exploiting the correlation information inside an analysis frame rather than inter-frame information. The performance of the proposed method starts to degrade when the hop size is larger than 4 and fails when the hop size is larger than 32.

Fig. 13(b) shows the computational time of the proposed method using different hop sizes. For the baseline method, whose performance is independent of the hop size, we only use a hop size of 1024, which has the smallest computational cost. Both algorithms were coded with Matlab and run on an Intel $i7@3.4GHz$ CPU with $16GB$ RAM. The largest computation time (using a hop size of 2) of the proposed method is still comparable to the baseline method with a hop size of 1024.

VI. Conclusions

We addressed the device self-localization problem in an ad-hoc sensor network by exploiting the TDOAs from multiple sound sources to asynchronous devices. We used the extreme (maximum and minimum) TDOAs from end-fire sound sources to calculate the relative distance between devices without knowing their time offsets. To estimate the extreme TDOAs, we proposed an audio-fingerprinting-based method, which extracts audio landmarks from noisy recordings and estimates the TDOA information by matching these landmarks. Using extracted landmark features consisting of pairs of spectral peaks of the audio signal for extreme TDOA estimation was found to be more robust to noise than the phase-based GCC algorithm.

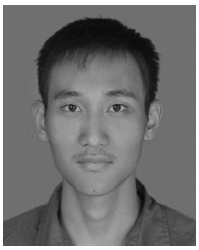
The proposed method assumes a sufficient number of sound sources around the ad-hoc array (i.e., end-fire sources). The deviation of the end-fire sources from their desired locations can lead to inter-device distance estimation errors. Performance improvement with non-ideal end-fire sources will be part of our future work.

References

[1] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing," *IEEE Commun. Mag.*, vol. 48, no. 9, pp. 140–150, Sep. 2010.
[2] M. Hennecke and G. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, Edinburgh, U.K., May 2011, pp. 127–132.
[3] H. Aghajan and A. Cavallaro, *Multi-camera networks: Principles and application*. New York, NY, USA: Academic, 2009.
[4] N. Drawil, H. Amar, and O. Basir, "GPS localization accuracy classification: A context-based approach," *IEEE Trans. Intell. Transportat. Syst.*, vol. 14, no. 1, pp. 262–273, Mar. 2013.
[5] N. Ravi, P. Shankar, A. Frankel, A. Elgammal, and L. Iftode, "Indoor localization using camera phones," in *Proc. IEEE Workshop Mobile Comput. Syst. Appl.*, Orcas Island, Aug. 2006, vol. Suppl., pp. 1–7.
[6] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: A high accuracy acoustic ranging system using cots mobile devices," in *Proc. ACM Int. Conf. Embedded Netw. Sensor Syst.*, Sydney, Australia, Nov. 2007, pp. 1–14.
[7] J. Schmalenstroeer, P. Jebramcik, and R. Haeb-Umbach, "A combined hardware-software approach for acoustic sensor network synchronization," *Signal Process.*, vol. 107, pp. 171–184, Feb. 2015.
[8] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Process.*, vol. 107, pp. 185–196, Feb. 2015.
[9] J. Bahi, A. Makhoul, and A. Mostefaoui, "A mobile beacon based approach for sensor network localization," in *Proc. IEEE Int. Conf. Wireless Mobile Comput., Netw., Commun.*, White Plains, NY, USA, Oct. 2007, p. 44.
[10] P. Pertila, M. Mieskolainen, and M. Hamalainen, "Passive self-localization of microphones using ambient sounds," in *Proc. Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 1314–1318.
[11] P. Pertila, M. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.
[12] A. L. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Conf. Music Info. Retrieval*, Baltimore, MD, USA, Oct. 2003, pp. 7–13.
[13] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[14] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in *Proc. ACM Int. Conf. Mobile Comput. Netw.*, Chicago, IL, USA, Sep. 2010, pp. 173–184.
[15] A. Tsui, W.-C. Lin, W.-J. Chen, P. Huang, and H.-H. Chu, "Accuracy performance analysis between war driving and war walking in metropolitan Wi-Fi localization," *IEEE Trans. Mobile Comput.*, vol. 9, no. 11, pp. 1551–1562, Nov. 2010.
[16] Z. Zhong and T. He, "MSP: Multi-sequence positioning of wireless sensor nodes," in *Proc. ACM Int. Conf. Embed. Netw. Sens. Syst.*, Sydney, Australia, Nov. 2007, pp. 15–28.
[17] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher, "Range-free localization schemes for large scale sensor networks," in *Proc. ACM Int. Conf. Mobile Comput. Netw.*, San Diego, CA, USA, Sep. 2003, pp. 81–95.
[18] D. Niculescu and B. Nath, "DV based positioning in ad hoc networks," *Tele. Syst.*, vol. 22, pp. 267–280, Jan. 2003.
[19] P. Robertson, M. Angermann, and B. Krach, "Simultaneous localization and mapping for pedestrians using only foot-mounted inertial sensors," in *Proc. ACM Int. Conf. Ubiquitous Comput.*, Orlando, FL, USA, Sep. 2009, pp. 93–96.
[20] O. Woodman and R. Harle, "Pedestrian localization for indoor environments," in *Proc. ACM Int. Conf. Ubiquitous Comput.*, Seoul, Korea, Sep. 2008, pp. 114–123.
[21] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 523–535, Apr. 2002.
[22] B. Williams, G. Klein, and I. Reid, "Real-time SLAM relocalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio, Brazil, Oct. 2007, pp. 1–8.
[23] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.
[24] M. Hebert, "Active and passive range sensing for robotics," in *Proc. IEEE Int. Conf. Roboti. Autom.*, San Francisco, CA, USA, 2000, vol. 1, pp. 102–110.
[25] N. Anjum and A. Cavallaro, "Automated localization of a camera network," *IEEE Intell. Syst.*, vol. 27, no. 5, pp. 10–18, Sep. 2012.
[26] R. Mohedano, A. Cavallaro, and N. Garcia, "Camera localization using trajectories and maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 684–697, Apr. 2014.
[27] M. Kushwaha, K. Molnar, J. Sallai, P. Volgyesi, M. Maroti, and A. Ledeczi, "Sensor node localization using mobile acoustic beacons," in *Proc. IEEE Int. Conf. Mobile Adhoc and Sens. Syst.*, Washington, DC, USA, Nov. 2005, pp. 1–9.
[28] M. Youssef, A. Youssef, C. Rieger, U. Shankar, and A. Agrawala, "Pin-Point: An asynchronous time-based location determination system," in *Proc. ACM Int. Conf. Mobile Syst., App. Services*, Uppsala, Sweden, Jun. 2006, pp. 165–176.
[29] M. Erol-Kantarci, H. Mouftah, and S. Oktug, "Localization techniques for underwater acoustic sensor networks," *IEEE Commun. Mag.*, vol. 48, no. 12, pp. 152–158, Dec. 2010.
[30] J. Herrera and H. S. Kim, "Ping-Pong: Using smartphones to measure distances and relative positions," in *Proc. 166 ASA Meet. Acoust.*, San Francisco, CA, USA, Dec. 2014, vol. 20, no. 1, pp. 1–10.
[31] W. Xi, Y. He, Y. Liu, J. Zhao, L. Mo, Z. Yang, J. Wang, and X. Li, "Locating sensors in the wild: Pursuit of ranging quality," in *Proc. ACM Conf. Embed. Netw. Sens. Syst.*, Zurich, Switzerland, Nov. 2010, pp. 295–308.
[32] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Push the limit of WiFi based localization for smartphones," in *Proc. ACM Int. Conf. Mobile Comput. Netw.*, Istanbul, Turkey, Aug. 2012, pp. 305–316.
[33] J. Qiu, D. Chu, X. Meng, and T. Moscibroda, "On the feasibility of real-time phone-to-phone 3D localization," in *Proc. ACM Conf. Embed. Netw. Sens. Syst.*, Seattle, WA, USA, Nov. 2011, pp. 190–203.
[34] H. Liu, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Accurate WiFi based localization for smartphones using peer assistance," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2199–2214, Oct. 2013.
[35] V. Filonenko, C. Cullen, and J. Carswell, "Investigating ultrasonic positioning on mobile phones," in *Proc. Int. Conf. Indoor Position. Indoor Nav*, Zurich, Switzerland, Sep. 2010, pp. 1–8.
[36] M. Hazas, C. Kray, H. Gellersen, H. Agbota, G. Kortuem, and A. Krohn, "A relative positioning system for co-located mobile devices," in *Proc. ACM Int. Conf. Mobile Syst., Appl., Services*, Seattle, WA, USA, Jun. 2005, pp. 177–190.
[37] D. Moore, J. Leonard, D. Rus, and S. Teller, "Robust distributed network localization with noisy range measurements," in *Proc. ACM Int. Conf. Embed. Netw. Sens. Syst*, Baltimore, MD, USA, Nov. 2004, pp. 50–61.
[38] S. Thrun, "Affine structure from sound," in *Adv. Neural Inf. Process. Syst.*, 2005, pp. 1353–1360.
[39] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, Feb. 2012.
[40] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 70–83, Jan. 2005.
[41] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New York, NY, USA, 2009, pp. 161–164.
[42] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 106–110.
[43] Y. Kuang, S. Burgess, A. Torstensson, and K. Astrom, "A complete characterization and solution to the microphone position self-calibration problem," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3875–3879.
[44] S. T. Birchfield and A. Subramanya, "Microphone array position calibration by basis-point classical multidimensional scaling," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1025–1034, Sep. 2005.
[45] J. H. Walters, R. S. Wilson, and J. S. Abel, "Speaker array calibration using inter-speaker range measurements," in *Proc. 116th AES Conv.*, Berlin, Germany, May 2004, no. 6038, pp. 1–8.
[46] R. S. Wilson, J. H. Walters, and J. S. Abel, "Speaker locations from inter-speaker range measurements: Closed-form estimator and performance relative to the Cramer-Rao lower bound," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, May 2004, vol. 2, pp. 389–392.
[47] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 666–670, Mar. 2008.
[48] M. J. Taghizadeh, P. N. Garner, and H. Bourlard, "Enhanced diffuse field model for ad hoc microphone array calibration," *Signal Process.*, vol. 101, pp. 242–255, Aug. 2014.
[49] N. Duong and F. Thudor, "Movie synchronization by audio landmark matching," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 3632–3636.
[50] N. Bryan, P. Smaragdis, and G. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 2389–2392.
[51] C. Cotton and D. Ellis, "Finding similar acoustic events using matching pursuit and locality-sensitive hashing," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2009, pp. 125–128.
[52] D. Ellis, Robust landmark-based audio fingerprinting 2009 [Online]. Available: http://labrosa.ee.columbia.edu/matlab/fingerprint/
[53] L. Kennedy and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," in *Proc. ACM Int. Conf. World Wide Web*, Madrid, Spain, Apr. 2009, pp. 311–320.
[54] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.
[55] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 21–24.
[56] K. D. Donohue, Audio systems array processing toolbox, 2009. [Online]. Available: http://www.engr.uky.edu/donohue/audio/Arrays/MA-Toolbox.htm

Tsz-Kin Hon received his B.Eng. degree in electronic and computer engineering from the Hong Kong University of Science and Technology (HKUST) in 2006; and the Ph.D. degree in digital signal processing from Kings College London (KCL) in 2013. He was a Research Engineer in the R&D of the Giant Electronic Ltd. between 2006 and 2009. He is currently a Postdoctoral Research Assistant in the Centre for Intelligent Sensing at Queen Mary University of London. His research interests include audio and video signal processing, device localization and synchronization, multi-source signal processing, joint time–frequency analysis and filtering, acoustic echo cancellation, speech enhancement, and biomedical signal processing.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2003 and the Ph.D. degree in signal processing from Dalian University of Technology, Dalian, China, in 2010 From 2011 to 2013, he was an Alexander von Humboldt Fellow in University of Oldenburg, Germany. Since 2014, he has been a Postdoctoral Researcher with the Centre for Intelligent Sensing, Queen Mary University of London, U.K. His research interests include video and audio compression, blind source separation, and 3D audio processing.

Joshua D. Reiss is a Reader in Audio Engineering with the Centre for Digital Music in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He has bachelors degrees in both physics and mathematics, and earned his Ph.D. in physics from the Georgia Institute of Technology. He is a member of the Board of Governors of the Audio Engineering Society, and co-founder of the company MixGenius, now known as LandR. Dr. Reiss has published more than 100 scientific papers and serves on several steering and technical committees. He has investigated sound synthesis, time scaling and pitch shifting, source separation, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.

Andrea Cavallaro received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He was a Research Fellow with British Telecommunications, London, U.K., from 2004 to 2005. He currently is a Professor of Multimedia Signal Processing and the Director of the Centre for Intelligent Sensing with Queen Mary University of London, London. He has authored more than 130 journal and conference papers, one monograph on Video Tracking (Wiley, 2011), and three edited books, *Multi-Camera Networks* (Elsevier, 2009), *Analysis, Retrieval and Delivery of Multimedia Content* (Springer, 2012), and *Intelligent Multimedia Surveillance* (Springer, 2013). Prof. Cavallaro is an Area Editor of *IEEE Signal Processing Magazine* and Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He is an elected member of the IEEE Signal Processing Society, and the Image, Video, and Multidimensional Signal Processing Technical Committee, and is the Chair of its Awards Committee. He served as an elected member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee, as an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON SIGNAL PROCESSING, and as a Guest Editor of seven international journals. He was the General Chair for the IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was a Technical Program Chair of the IEEE AVSS 2011, the European Signal Processing Conference in 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards on target tracking and perceptually sensitive coding at the IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.