A Survey on Acoustic Sensing

Chao Cai, ACM Student Member
Rong Zheng, ACM Member
Menglan Hu, ACM Member

The rise of Internet-of-Things (IoT) has brought many new sensing mechanisms. Among these mechanisms, acoustic sensing attracts much attention in recent years. Acoustic sensing exploits acoustic sensors beyond their primary uses, namely recording and playing, to enable interesting applications and new user experience. In this paper, we present the first survey of recent advances in acoustic sensing using commodity hardware. We propose a general framework that categorizes main building blocks of acoustic sensing systems. This framework consists of three layers, i.e., the physical layer, processing layer, and application layer. We highlight different sensing approaches in the processing layer and fundamental design considerations in the physical layer. Many existing and potential applications including context-aware applications, human-computer interface, and aerial acoustic communications are presented in depth. Challenges and future research trends are also discussed.

Additional Key Words and Phrases: Acoustic sensing, aerial acoustic communication, context-aware applications, human-computer interface.

1. INTRODUCTION

Internet of Things (IoT) [Yang et al. 2017b] technologies enable everyday objects to connect and communicate with each other by augmenting them with sensing, processing, and computation units. With the ever increasing computation power and rich built-in sensors available for IoT devices, new sensing methodologies are emerging that repurpose sensors beyond its primary use. For instance, cameras are intended for taking photos but have been utilized in visible light communication [Yang et al. 2017a]. Gyroscope and accelerometer sensors are designed for attitude estimation but have been used extensively in activity recognition [Liu et al. 2016]. WiFi signals, originally used for communication, have been widely applied in context-aware computing [Pu et al. 2013; Adib et al. 2014; Kumar et al. 2014; Vasisht et al. 2016]. In this paper, we target innovative sensing mechanisms that exploit acoustic sensors.

Acoustic sensors, namely microphones and speakers, are one of the most commonly used transducers in IoT devices. They are generally used for playing and recording audio signals and have already played a pivotal role in a myriad of applications such as speech recognition [Boll 1979; Sakoe and Chiba 1978], audio beamforming, and source localization [Chen et al. 2002; Rafaely 2005]. Nowadays smart IoT products with acoustic sensors and cloud-based machine learning technologies are gaining popularity. Examples are Google Home [CNET 2018b] and Amazon Echo [CNET 2018a]. However, these developments are limited to passive acoustic sensing in the human audible frequency range, and thus leave many untapped potentials to be explored.

Novel sensing mechanisms are emerging that treat acoustic sensors as transceivers that emit and capture wireless signals. For instance, acoustic signals have been used to establish aerial acoustic communication channels to transmit a small amount of information [Ka et al. 2016; Nandakumar et al. 2013; Wang et al. 2016]. Like Radio Frequency (RF) signals, acoustic signals can be reflected by obstacles, which allows the development of acoustic-enabled short-range radars for floor map reconstruc-

Author's addresses: C. Cai and M. Hu are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, China; R. Zheng is with the School of Computing and Software, McMaster University, Canada.

DOI: http://dx.doi.org/10.1145/0000000.0000000

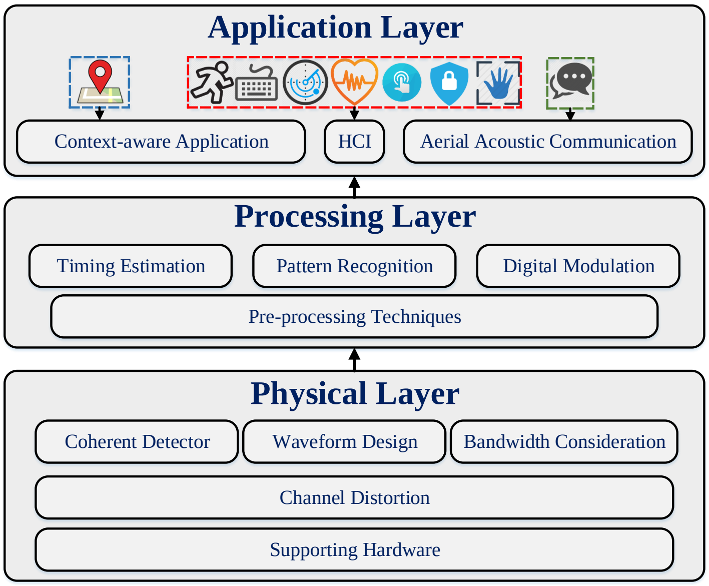


Fig. 1: A general framework for acoustic sensing

tion [Zhou et al. 2017] and gesture recognition [Nandakumar et al. 2016; Wang et al. 2016; Yun et al. 2017]. The relatively slow propagation speed of acoustic waves in common medium (compared to RF) makes it possible to achieve accurate Time-of-Flight estimation, enabling many context-aware applications [Peng et al. 2007; Uddin and Nadeem 2013; Lazik et al. 2015; Lazik and Rowe 2012; Liu et al. 2012; Liu et al. 2013; Wang et al. 2017; Wang et al. 2016]. For acoustic emitting sources, their unique signatures can be utilized for authentication or activity recognition [Chauhan et al. 2017]. For non-acoustic emitting objects, active sensing can be adopted, which transmits purposefully modulated acoustic signals and makes inference based on the received or reflected waveforms at an acoustic receiver [Gupta et al. 2012; Ke et al. 2018]. [1]

Despite tremendous development in acoustic sensing techniques in the past decade, a comprehensive treatment of key design considerations, fundamental principles, and methods are lacking. As a result, when developing a new application utilizing acoustic sensing, researchers and developers often have to start from scratch and reinvent the wheel.

In this work, we present the first survey on recent advances in acoustic sensing. We target novel sensing approaches on commodity hardware (bandwidth below 24 kHz), as opposed to those that require special-purposed hardware such as underwater acoustic communication or ultrasonic sensing. We survey the state-of-the-art research and propose a general framework that encompasses main building blocks of typical acoustic sensing systems. This framework provides a novel layered taxonomy of previous work consisting of application layer, processing layer, and physical layer as depicted in Fig. 1. In the framework, the physical layer converts acoustic signals into digital samples with appropriate hardware and signal processing modules; the processing layer extracts application-specific information such as unique temporal, frequency-domain,

[1] It should be noted that communication and sensing are different terminologies. However, in this paper, to fully uncover the potential of acoustic signals, we view acoustic communication as a special type of acoustic sensing, whose purpose is to convey information.

Table I: Comparison of acoustic sensing enabled applications

Category	**Application**	**Work**	**Sensing method**	**Platform**
Context-aware application	Ranging	[Peng et al. 2007] [Uddin and Nadeem 2013]	Active sensing	Mobile devices
	Radar	[Graham et al. 2015] [Zhou et al. 2017]		
	Localization	[Lazik et al. 2015] [Lazik and Rowe 2012] [Liu et al. 2013] [Wang et al. 2017] [Tung and Shin 2015]		
	Device-based tracking	[Mao et al. 2016] [Mao et al. 2017] [Yun et al. 2015]		
HCI	Device-free gesture tracking	[Nandakumar et al. 2016] [Wang et al. 2016] [Yun et al. 2017]	Passive or active sensing	Mobile devices or customized hardware
	Keystroke detection	[Liu et al. 2015] [Wang et al. 2014] [Zhu et al. 2014]		
	Acoustic-driven interface	[Laput et al. 2015] [Ono et al. 2015]		
	Touch force detection	[Ono et al. 2015] [Pedersen and Hornbæk 2014] [Tung and Shin 2016]		
	Gesture recognition	[Gupta et al. 2012] [Ruan et al. 2016]		
	Health sensing	[Larson et al. 2012] [Larson et al. 2011] [Lu et al. 2012] [Nandakumar et al. 2015] [Nirjon et al. 2012] [Rachuri et al. 2010] [Ren et al. 2015]		
	Authentication	[Chauhan et al. 2017]		
	Activity recognition	[Nirjon et al. 2013] [Rahman et al. 2014] [Zhang et al. 2016]		
Aerial acoustic communication	Communication	[Ka et al. 2016] [Lee et al. 2015] [Nandakumar et al. 2013] [Wang et al. 2016]	Active sensing	Mobile devices

or high-level features; the application layer leverages the information from the processing layer and provides a variety of services to end-users.

The remainder of this paper is organized as follows: in Section 2, we classify existing work according to their application scenarios into three categories, namely, context-aware applications, human-computer interface (HCI), and aerial acoustic communication. In Section 3, we summarize the building blocks for the above applications and discuss key enabling techniques including timing estimation, pattern recognition, and digital modulation. In Section 4, we present supporting hardware and physical layer design considerations, including coherent detector, waveform design, and bandwidth consideration. Finally, we discuss remaining challenges and new research directions in Section 5. We conclude the paper in Section 6.

2. APPLICATION LAYER

In this section, we discuss various acoustic-enabled applications. Based on the application scenarios, we classify existing research into three categories: *context-aware application*, *HCI*, and *aerial acoustic communication*. Different categories exploit acoustic signals in different manners. Context-aware applications, depending on contextual information such as range, location, etc., rely on the estimation of sound propagation time. HCI systems infer and respond to user intentions by inspecting how external physical activities alter acoustic signals. Aerial acoustic communication utilizes acoustic waves to carry data through air. These applications mostly adopt *active* sensing where modulated acoustic waves are generated. A comparison of these applications in sensing methodologies, occupied bandwidth, and deployed platforms is summarized in Table I.

2.1. Context-aware Applications

Context-aware applications, built on contextual information such as range and location can provide better user experience in many domains such as health and fitness, entertainment, etc. The context-aware applications, depending on the estimation of sound propagation time or special acoustic signatures, can be further grouped into four categories: ranging, acoustic radar, device-based tracking, and localization.

2.1.1. Ranging. Range is a useful context information and can be used for distance and size measurements, efficient network management [Nadeem and Ji 2007], or content sharing [Frohlich et al. 2002; Counts and Fellheimer 2004]. Leveraging acoustic signals for ranging provides an economical and convenient alternative to traditional measurement tools.

BeepBeep [Peng et al. 2007] is a pioneer work that uses acoustic signal for precise ranging on commodity mobile devices. It calculates the distance between pair-wise transceivers by estimating the propagated time of acoustic signals. BeepBeep avoids tight synchronization via a two-way sensing approach. In BeepBeep, one device first emits a chirp signal. Upon detection, the other device waits for an arbitrary period and then emits another chirp signal. Both transceivers then calculate the time difference between the events of transmission and reception locally by counting the number of acoustic samples. A central server is used to compute the final results from the time differences. BeepBeep reports centimeter-level ranging accuracy. However, the performance of BeepBeep can be degraded by irregular and uncertain system delay. To mitigate such uncertainty, the authors in [Uddin and Nadeem 2013] sidestep the system jitters by implementing the system in the kernel space, and build a stand-alone application called RFBeep. RFBeep [Uddin and Nadeem 2013] performs ranging via a combination of radio and acoustic signals. The key idea for RFBeep is that the flight time of radio signals is negligible in the maximum reachable distance for power-limited acoustic signals. Thus it is feasible to employ radio signals for synchronization. As a result, the range information can be acquired via estimating the flight time of acoustic signals. RFBeep reports 20 cm absolute ranging errors. However, the reliance on kernel modification prohibits its wide-scale adoption. SwordFight [Zhang et al. 2012] is another ranging system that improves upon BeepBeep in responsiveness, accuracy, and robustness. It works at the same way as BeepBeep but differs in employed signals and signal detection techniques. SwordFight reports a median ranging accuracy of 2 cm with 10 Hz fresh rate in noisy environments.

2.1.2. Acoustic Radar. Radar [Philippe et al. 2001] is widely used for remote sensing [Cloude and Pottier 1997] and object tracking [Blackman 1986; Cloude and Pottier 1996]. A radar system operates by radiating high-frequency signals and detecting

Table II: Comparison of acoustic-enabled localization systems

Category	Work	Method	Synchronous/ asynchronous	Concurrent multiple targets localization	Accuracy
Infrastructure-based	[Liu et al. 2013]	ToA	Synchronous	Supported	Centimeter-level
	[Lazik and Rowe 2012]	TDoA	Synchronous	Supported	Centimeter-level
	[Lazik et al. 2015]	TDoA	Synchronous	Supported	Centimeter-level
	[Wang et al. 2017]	TDoA	Asynchronous	Supported	Centimeter-level
Infrastructure-free	[Liu et al. 2012]	ToA	Asynchronous	Not supported	Meter-level
	[Nandakumar et al. 2012]	TDoA	Asynchronous	Supported	Meter-level
	[Tung and Shin 2015]	Profiling	N/A	Not Supported	Centimeter-level

the reflected echoes [ztrk et al. 2017; Philippe et al. 2001]. BatMapper [Zhou et al. 2017] demonstrates that a radar can be realized using acoustic signals on off-the-shelf mobile devices for indoor floor map construction. BatMapper adopts the FMCW (Frequency Modulated Continuous Wave) signals and exploits the constraints of speaker-microphone distances to detect the echoes bouncing off surrounding objects, leading to accurate range estimation. BatMapper reports $1-2$ cm estimation errors with ranges up to 4 meters. Another work similar to BatMapper is presented in [Graham et al. 2015], reporting an error bound of 12 cm within 4 m distances.

2.1.3. Device-based Tracking. High-accuracy object tracking is important in many scenarios [Yilmaz et al. 2006], for example, interaction, automated surveillance, and traffic monitoring. Tracking is already a well-investigated topic in Computer Vision (CV) community [Yilmaz et al. 2006; Murray 2017; Xiang et al. 2015]. However, CV techniques impose substantial computation costs and do not work well under poor light conditions. Acoustic-based tracking systems can overcome these limitations.

AAMouse [Yun et al. 2015] is a device-based tracking system. It achieves centimeter-level accuracy by estimating the Doppler effect of multiple acoustic carriers. However, the sampling rate limits the tracking accuracy, and the tracking errors accumulate over time, making it infeasible for long-term tracking. CAT [Mao et al. 2016], advances AAMouse and achieves a sub-centimeter level tracing accuracy by a chirp mixing operation. CAT also incorporates an Inertial Measurement Unit (IMU) to improve tracking accuracy. However, CAT adopts one-way sensing which is easily affected by the Sampling Frequency Offset (SFO) [Kinoshita and Nakatani 2013; Miyabe et al. 2013a; Miyabe et al. 2013b]. The SFO problem exhibits irregularity and cannot be addressed by a one-time compensation. Therefore, CAT needs to perform calibration from time to time. The authors further advance CAT in [Mao et al. 2017], enabling a drone to follow a person with a safe range in challenging indoor environments. In this work, the authors utilize several advanced signal processing modules, in particular, MUltiple SIgnal Classification algorithm (MUSIC) to resolve the multipath effects and thus enhance the robustness of tracking.

2.1.4. Localization. Localization is the key enabler for Location Based Service (LBS). Though there is tremendous research on indoor localization, they either use expensive dedicated infrastructures [Yang et al. 2014; Adib et al. 2014; Xiong and Jamieson 2013] or rely on cumbersome device-dependent kernel modification [Vasisht et al. 2016; Kumar et al. 2014; Kotaru et al. 2015], prohibiting their practical deployment. Decades of efforts have been made yet indoor localization services are still not widely available. Among existing cutting-edge indoor localization approaches, acoustic-based systems attract much interest in the community since they can achieve sub-meter level localization accuracy with relatively low infrastructure cost and deployment efforts.

Existing work on acoustic-enabled localization solutions can be classified into two categories, namely, *infrastructure-based* and *infrastructure-free*.

Infrastructure-based schemes often deploy low-cost and power-efficient distributed acoustic anchors in the place-of-interest. The coordinates of these anchors are measured in advance. Apart from acoustic sensors, each anchor may have a wireless connection with a remote server. The remote server synchronizes or schedules the anchors in transmitting modulated signals. When these signals are detected by either a target or other anchors, the associated timestamps (time-of-arrival or time-difference-of-arrival) are reported to the sever. Finally, the location of a target is obtained. In contrast, infrastructure-free systems require no extra hardware but usually sacrifice localization accuracy. A comparison of acoustic-enabled indoor localization systems is summarized in Table II.

Liu et al. [Liu et al. 2013] developed a centimeter-level localization system named Guoguo. The anchors in this system are synchronized by Zigbee and are scheduled to transmit orthogonal codes, which are used by targets to perform ToA (Time-Of-Arrival) estimation. Multilateration is then used to locate the targets. A speaker-only localization system was proposed by Lazik and Rowe in [Lazik and Rowe 2012]. In this approach, distributed speakers are connected to different channels of an advanced audio device. Each channel emits chirp spread spectrum modulated acoustic signals [Kim and Chong 2015]. A target locates itself locally by performing TDoA (Time-Difference-Of-Arrival) estimation. According to [Lazik and Rowe 2012], the 95-percentile localization accuracy is within 10 cm. ALPS [Lazik et al. 2015] improves upon the work [Lazik and Rowe 2012] in ease of deployment. In ALPS, anchors are synchronized via Bluetooth, and each anchor embeds both a microphone and a speaker. The coordinates of the anchors are efficiently obtained through acoustic-assisted simultaneously localization and mapping. ALPS reports average errors of 30 cm and 16.1 cm in locating targets and anchors. The localization accuracy of the above work is highly dependent on the synchronization performance, which is sensitive to network latency, especially in a large-scale network. In contrast, asynchronous approaches can overcome these shortcomings. ARABIS [Wang et al. 2017] is an asynchronous acoustic localization system which adopts two-way ranging [Peng et al. 2007] to avoid synchronization. In ARABIS, anchors transmit acoustic beacons periodically following a coarse time-division-multiple-access schedule. Targets, as well as anchors, overhear the transmissions and record the corresponding timestamps. These timestamps can be used to estimate TDoA information in locating a target. ARABIS reports a 95-percentile localization error of 7.4 cm.

Infrastructure-free localization systems do not require the deployment of custom-built hardware in the place-of-interest but achieve less competitive localization accuracy. Liu et al. in [Liu et al. 2012] built a localization system utilizing acoustic and WiFi signal. This approach first obtains pair-wise distances between different mobile devices via acoustic ranging [Peng et al. 2007]. The ranging results are then used to bound the spatial relations in the device group to form a rigid graph. Furthermore, the device group collect location-dependent WiFi signatures. The system finally locates the device group by finding a best match for the spatial constraints and location-dependent WiFi signatures. This work achieves an 80-percentile localization error of 1 m. Centaur [Nandakumar et al. 2012], similar to [Liu et al. 2012], proposes a joint optimization framework utilizing acoustic and WiFi signals, and reports meter-level localization accuracy. EchoTag [Tung and Shin 2015] is an acoustic-based fingerprinting localization system that can detect minor location changes. It associates different acoustic profiles with different positions, known as tags, to train a classification model. This model is then used for online tag detection and enable context-aware applications. EchoTag reports an accuracy of 98% in distinguishing 11 tags at 1 cm resolution. How-

Table III: A Comparison of device-free gesture tracking systems

System	Transmitted signal	Average tracking accuracy	Occupied bandwidth	Fresh rate	Operation range
FingerIO	OFDM modulated signal	8 mm in 2D	$18 - 20$ kHz	169 fps	within 0.5 m
LLAP	Multiple pure tones	3.5 mm for 1D 4.57 mm for 2D	$17 - 23$ kHz	≥ 66 fps	0.5 m
Strata	GSM sequence	3 mm	$18 - 22$ kHz	80 fps	0.5 m

ever, EchoTag cannot adapt to environmental dynamics and is likely to suffer from degraded performance in the long run in absence of new data collections.

We believe that acoustic-enabled infrastructure-based localization systems are promising for commercial adoption since they can achieve high localization accuracy with affordable infrastructure costs and can be deployed on commodity mobile devices. Infrastructure-free solutions can be supplementary to the infrastructure-based systems in cases where anchor nodes cease to operate and do not have sufficient coverage of areas of interest.

2.2. HCI

HCI [Lowgren 2014], a multidisciplinary field of study, focuses on information technology design, in particular, the interaction between humans and computers. Acoustic-enabled HCI is an emerging modality that exploits the relation between acoustic channel properties and physical activities. Such a technology can be applied to a wide range of interactions from finger-scale tracking to body-scale activity recognition. In this section, we group existing work that utilizes acoustic-based HCI methods into eight categories, namely, device-free gesture tracking, keystroke detection, acoustic-driven interface, touch force detection, gesture recognition, audio-based health sensing, authentication, and activity recognition.

2.2.1. Device-free Gesture Tracking. Device-free gesture tracking, a type of Around Device Interaction [Ketabdar et al. 2010] (ADI), extends interaction space beyond the physical boundary of small mobile devices and effectively uses the nearby 3D space for interaction. This kind of technology is particularly useful for small portable devices such as smartphones and wearables.

FingerIO [Nandakumar et al. 2016] is a recently proposed device-free gesture tracking system. It turns a mobile phone or a smartwatch to an active sonar that is capable of tracking moving fingers at a granularity of 7 mm. FingerIO uses the Orthogonal Frequency Division Multiplexing (OFDM) modulated signals to estimate the acoustic channel between a hand and the smartphone periodically. In each estimation cycle, the channel responses, also called channel frames, are acquired through cross-correlation. Since only moving objects can dynamically affect the channel, FingerIO can track a moving object by comparing consecutive channel frames. FingerIO reports 8 mm median tracking accuracy at a frame rate of 169 fps. A phase-based device-free gesture tracking system, LLAP, was proposed in [Wang et al. 2016]. LLAP leverages coherent detection to extract the phase of the acoustic echoes for finger localization and tracking. In LLAP, a mobile device actively transmits multiple carriers and decomposes finger-generated echoes for processing. It uses the phase divergence of multiple carriers to coarsely locate the finger and track its displacement via phase shift. LLAP reports a tracking accuracy of 3.5 mm for 1D hand movement and 4.57 mm for 2D drawing with less than 15 ms latency. However, both FingerIO and LLAP are not resilient to nearby interference. Another work named Strata was proposed in [Yun et al. 2017]. Strata is also built on the coherent detector structure but utilizes GSM training sequence. The evaluation results demonstrate that it outperforms FingerIO and LLAP in all cases. A comparison of these tracking schemes is given in Table III.

Device-free gesture tracking systems enable drawing in-the-air experiences on tangible devices. However, these systems still have a long way to go before they reach massive market. Initial setups, multipath effects, and the placement of acoustic sensors all affect tracking performance. As a result, more robust approaches that are adaptive to different settings need to be investigated.

2.2.2. Keystroke Detection. Keystroke detection via acoustic sensing can provide an alternative input method for the current inefficient and error-prone touchscreen keyboards. Such a technology can also be used by malicious attacks to hack sensitive information.

UbiK [Wang et al. 2014] is a novel keystroke recognition system based on passive acoustic sensing. It uses a printed paper to emulate a keyboard, enabling PC-like text input. UbiK harnesses the fact that the amplitude spectrum density of acoustic signals, produced by click sound, is location dependent. Thus a fingerprinting strategy can be employed for keystroke detection. UbiK reports 95-percentile recognition accuracy. However, UbiK needs laborious training. Model-based methods can ease the pain of laborious training. The authors in [Liu et al. 2015] and [Zhu et al. 2014] apply TDoA inference models to perform keystroke detection. They use the dual microphones on smartphones to passively locate the pressed keys by correlating the associated audio samples from the two channels. The work in [Liu et al. 2015] and [Zhu et al. 2014] report 94% and 72.2% detection accuracy, respectively.

Keystroke detection via acoustic sensing has demonstrated both analytically and empirically to be viable. However, heterogenous typing styles across different performers, background noises, and different typing surfaces, etc., may deteriorate system performance. Therefore, there is still room for further improvements.

2.2.3. Acoustic-driven Interface. Building acoustic-driven interface with custom-built hardware enjoys more design flexibility and can lead to brand new experiences. We now present novel interactive systems on customized platforms.

Touch & Active [Ono et al. 2015] is an active sensing system that can recognize a rich context of touch gestures and hand postures on existing objects. It can also identify different shapes of deformable objects. Touch & Active exploits the facts that any external excitations including touch or deformative force can alter the resonant frequency spectra of a specific object. Thus, if we use acoustic signals to monitor the properties of the object, external forces can be identified. Touch & Active reports an accuracy of 99.6% and 86.3% in recognizing five touch gestures and six hand postures on a plastic toy. Another work called Acoustruments [Laput et al. 2015] also adopts customized platforms. Acoustruments fabricated a tube as an acoustic channel that connects the transceivers on a commodity smartphone. The tube has a physical control unit that can move and thus manipulates the properties of received acoustic signals. Acoustruments designed a classification model to recognize different control commands and achieves 99% control accuracy.

2.2.4. Touch Force Detection. HCI with touchable panels can enhance input flexibility. However, prototyping force-sensitive input systems often require complex circuit designs and hardware configurations. Commodity hardware, e.g., smartphone, with such capability are based on priority sensors [3DT 2017] which are not pervasively adopted. In contrast, acoustic sensing modules are ubiquitous and can enable touch sensing on off-the-shelf smart devices without extra hardware.

ForcePhone [Tung and Shin 2016] can estimate applied forces with built-in acoustic sensors by exploiting structure-borne sound propagation. ForcePhone utilizes the fact

Table IV: A comparison of aerial acoustic communication systems

Work	Modulation	Maximum operating range	Bandwidth	Audible/ inaudible	Bit rate
[Gerasimov and Bender 2000]	OFDM	< 2 m	$735 - 4410$ Hz/ 18.4 Hz	Audible/ inaudible	5.6/ 1.4 kbps
[Lopes and Aguiar 2001]	M-ary FSK	< 2 m	$0 - 12$ kHz	Audible	2.4 kbps
[Hanspach and Goetz 2014]	FHSS	20 m	$4.1 - 21$ kHz	Audible	20 bps
[Nandakumar et al. 2013]	OFDM	Within centimeters	$0 - 24$ kHz	Audible	2.4 kbps
[Lopes and Aguiar 2006]	OFDM	8 m	$6.4 - 8$ kHz	Inaudible	240 bps
[Yun et al. 2010]	Phase modulation	< 2 m	$6.4 - 8$ kHz	Inaudible	600 bps
[Wang et al. 2016]	OFDM	8 m	$8 - 20$ kHz	Inaudible	500 bps
[Lee et al. 2015]	BOK	25 m	$19.5 - 22$ kHz	Inaudible	16 bps
[Ka et al. 2016]	QOK	2.7 m at 35 dBSPL	$18.5 - 19.5$ kHz	Inaudible	15 bps

that emitted acoustic signals from the speaker of a smartphone can cause vibration of the phone body, the intensity of which can be affected by external pressure. Therefore, touch force can be obtained by correlating the applied force with received signal intensity. ForcePhone reports comparable performance with iPhone 6s devices which feature $3D$ touch sensors. Touch & Active [Ono et al. 2015] which exploits the relationship between touch events and channel responses can also recognize applied forces. Since Touch & Active adopts classification model, it can only identify discrete forces. Expressive Touch [Pedersen and Hornbæk 2014] is a passive touch-sensitive system. It leverages the differences in signal intensity and spectrum of original touch generated sounds to identify different forces. Expressive Touch achieves less competing results as it can only identify two levels of touch force.

Acoustic-enabled touch-force detection enriches input flexibility. Nevertheless, existing solutions are still immature for practical deployment due to laborious calibrations. More efforts should be made to ease such pain.

2.2.5. Gesture Recognition. Gesture recognition aims to understand the expressive meaning of body parts, serving as an interface for humans to interact with smart devices. Previous approaches [Mitra and Acharya 2007] often rely on dedicated devices or computational intensive image processing. In contrast, acoustic sensing methods are lightweight and can detect minor finger-scale gestures.

AudioGest [Ruan et al. 2016] is a hand gesture recognition system with an accuracy of 96%. AudioGest harnesses the fact that different hand gestures generate different acoustic echo profiles. Therefore, one can construct sufficient gesture associated profiles to learn a classification model for online recognition. SoundWave, a Doppler effect based gesture recognition system, was presented in [Gupta et al. 2012]. SoundWave continuously triggers an inaudible tone and infers gestures by sensing the spectrum of hand-reflected echoes. The key idea is that the reflected acoustic echoes from a moving hand are frequency-shifted compared to the transmitted signals. If the hand is moving away, the spectrum of the acoustic echoes is below the transmitted one and vice versa. Combining the above primitives allows the recognition of more complex gestures such as flick and quick taps. SoundWave reports recognition accuracy over 86.67% under various testbeds. Other gesture recognition systems include VSin [Ke et al. 2018] that enables back-of-device gesture recognition and the work in [Sun et al. 2018] that facilitates depth-aware finger tapping on virtual displays. This line of work is based on gesture tracking technology.

Gesture recognition based on acoustic sensing mostly needs a sophisticated analysis on acoustic echoes. These echoes are submerged in primary signals and thus are hard to be isolated. Meanwhile, they are vulnerable to noise. As a result, acoustic-enabled gesture recognition systems have limited capability. Further investigation into signal process techniques and robust inference models are needed.

2.2.6. Audio-Based Health Sensing. Vital signs, including breathing rates and heartbeats, are important indicators of human health condition. Vital sign detection is previously limited to clinic usage as it requires special equipment and trained technicians. Acoustic sensing makes it feasible to detect vital sign signals with portable devices for non-professional end-users.

MusicalHeart [Nirjon et al. 2012] is a convenient, non-invasive, and low-cost smart device that recommends appropriate music to end-users based on their heartbeat rates. It exploited a customized hardware platform called Septimu to extract heartbeat signals from a resonant chamber inside a ear. Septimu achieves an average detection error of 7.5 bpm. Another portable life sign detection system based on commodity smartphones was presented in [Nandakumar et al. 2015]. This work leverages a smartphone as an active sonar to detect chest movements, enabling breathing rate estimation and sleep apnea detection. The proposed system can achieve an error of fewer than 0.11 breaths per minute even at a distance of up to 1 m. Ren et al. [Ren et al. 2015] developed a passive sensing system that can detect breathing rates and sleep-related events from the breathing signals. This approach employs more high-quality sensors, and reports less than 0.5 bpm detection error rate.

Other audio-based health sensing systems include the work in [Larson et al. 2011] that detects coughs, SpiroSmart [Larson et al. 2012] and SpiroCall [Goel et al. 2016] that diagnose lung function, EmotionSense [Rachuri et al. 2010] that identifies psychological information, and StressSense [Lu et al. 2012] that uncovers stress.

2.2.7. Authentication. Authentication via acoustic footprints is a well-investigated topic in the community. Traditional methods mostly utilize voice [Shoup et al. 2016] which require users' active involvements. However, in [Chauhan et al. 2017], Chauhan et al. [Chauhan et al. 2017] pioneered a novel passive method that utilizes behavioral biometric signatures for authentication. These signatures are extracted from a user's commonplace breathing signals including sniff, normal breaths, and deep breaths. These commonplace breathing signals exhibit distinctive features across different users and thus can be used for authentication. The prototype system in this work, called BreathPrint, reports an accuracy of over 94% in identifying different users. This approach is light-weight and promising for resource-constrained IoT devices.

2.2.8. Activity Recognition. Activity recognition is a well-investigated topic in the research community. It aims to understand the expressive meaning of human activities and react to the corresponding initiatives. Various methods have been put forwarded in the literature such as Inertial Measurement Unit (IMU) [Abhayasinghe and Murray 2014; Prathivadi et al. 2014; Koskimäki et al. 2017; Mummadi et al. 2017; Wei et al. 2016] assisted methods, RF signal [Wang et al. 2015; Virmani and Shahzad 2017; Abdelnasser et al. 2015; Li et al. 2016; Pu et al. 2013] aided approaches, and Computer Vision (CV) [Ma et al. 2016; Fernando et al. 2015; Pirsiavash and Ramanan 2012; Wang et al. 2013; Chéron et al. 2015] supported techniques. Leveraging acoustic signals for activity recognition provides an additional modality to improve accuracy.

DopEnc [Zhang et al. 2016] is an automatic encounter profiling system. It enables users to record conversation events and interaction contexts with other people automatically. The underlying techniques behind DopEnc are acoustic Doppler effect estimation and self-voice recognition. DopEnc achieves an accuracy of 6.9% false positive and 9.7% false negative rates in real-world usage. BodyBeat [Rahman et al. 2014] is a mobile sensing system that aims to recognize non-speech body sounds such as food intake, laughter, and breath. This approach requires users to wear a dedicated device around the neck to capture audio samples. It builds a classification model that employs 30 acoustic features for identification. BodyBeat is assumed to be useful for food journaling and illness detection. Auditeur [Nirjon et al. 2013] is a general-purpose acoustic event detection platform. It utilizes participatory sensing where end users tag audio clips for profiling. Auditeur improves the detection accuracy for acoustic events by 10.71% – 13.87% than traditional methods with 11.04% – 441.42% less power.

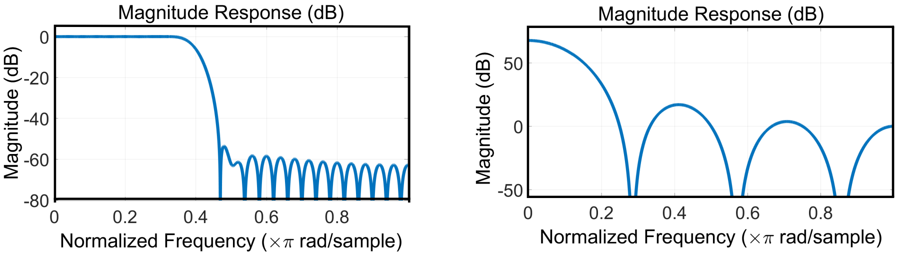
2.3. Aerial Acoustic Communication

With the advancement of mobile computing technologies, transmitting small amounts of data via aerial acoustic channels has attracted much attention, leading to a new concept of aerial acoustic communication. Aerial acoustic communication enables any device that has an embedded microphone and speaker to achieve communication without extra hardware and complex network configuration. Thus, it can serve as an alternative to traditional RF-based device-to-device communication such as Bluetooth and WiFi Direct.

The authors in [Gerasimov and Bender 2000] presented a communication system based on tone modulation. It leverages the presence or absence of tone signals to embed information (100% Amplitude Shift Keying). This approach achieves a delivery rate of 5.6 kbps with multiple audible tones. The delivery rate reduces to 1.4 kbps when a single inaudible tone is applied. It reaches a maximum communication range of 2 m under the Line-of-Sight (LOS) condition. Another work called Digital voice [Lopes and Aguiar 2001] adopted a M-ary FSK modulation mechanism with the audible band (under 12 kHz), reporting a data rate at tens to thousands of bits per seconds (bps). Dhwani [Nandakumar et al. 2013] is an acoustic-based Near Field Communication (NFC) system. It employs OFDM modulation to encode messages and design a Jam-Secure technique to prevent malicious attacking. Dhwani occupies 24 kHz bandwidth and achieves a maximum data rate of 2.4 kbps. An acoustic-enable mesh network was proposed in [Hanspach and Goetz 2014]. This approach leverages Frequency Hopping Spread Spectrum (FHSS) and achieves 20 bps at a distance up to 20 m. The above work all deploy systems in the audible band (normally below 18 kHz).

Utilizing audible bands for communication can be disruptive, and thus many inaudible (hidden) communication systems are developed. The authors in [Lopes and Aguiar 2006] proposed to leverage the masking effect of the human hearing system to achieve inaudible acoustic communication. This approach employs OFDM modulation and achieves 240 bps data rate. Similar work in [Yun et al. 2010] and [Wang et al. 2016] attains data rates of 600 and 500 bps, respectively.

Both tone-based and OFDM modulation techniques are not robust to Doppler effect. Besides, the performance of these methods generally deteriorates in multipath rich environments. In contrast, chirp spread spectrum (CSS) utilizes more resilient chirp signals to encode information bits and thus achieves better performance in decoding error rate and communication range. A chirp binary orthogonal keying (BOK) modulation techniques was first presented in [Berni and Gregg 1973; El-Khamy et al. 1996]. It utilizes orthogonal up and down chirp signals for communication. The work in [Lee et al. 2015] adopted BOK and extended the communication range up to 25 m at a data rate of 16 bps. Soonwon et al. [Ka et al. 2016] advanced BOK and developed a chirp quaternary orthogonal keying (QOK) modulation technique. QOK finds near-orthogonal chirps by an exhaustive search over its pre-defined solution space. With QOK, a Code Division Multiple Access (CDMA) system is built. The system achieves



(a) Frequency response of a FIR filter

(b) Frequency response of a CIC filter

Fig. 2: Frequency response of digital filters

zero frame error rate even at a minimal sound pressure level of 35 dB SPL when the transceivers are 2.7 m away from each other.

3. PROCESSING LAYER

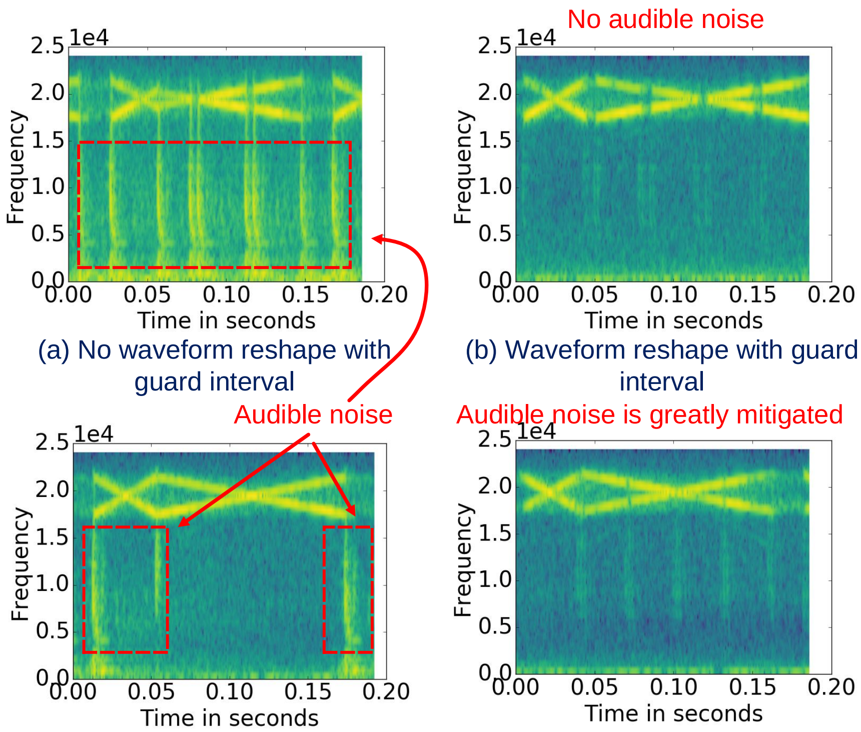
The processing layer serves as an intermediary between the physical and application layer. It takes audio samples from the physical layer, applies inference models to extract application-specific features, and provides the results to the application layer. How to undermine useful information via the inference models is central to the processing layer. In this section, we categorize existing approaches into *timing estimation*, *pattern recognition*, and *digital modulation*. Key techniques behind each category are presented in details. In timing estimation, different processing techniques to estimate sound propagation time are introduced; A canonical data flow to inspect data regularities is presented for pattern recognition. Common techniques to achieve aerial acoustic communication are compared in digital modulation. Before diving into the details of each category, we first present common pre-processing techniques among the three categories.

3.1. Pre-processing Techniques

Pre-processing techniques aim to achieve high signal-to-noise-ratio (SNR) since acoustic sensors especially microphones are quite sensitive and vulnerable to background noises, channel distortions, and multipath effect. In this section, we present the most widely used pre-processing techniques along the lines of noise filtering, channel distortion mitigation, and robust onset detection. The technique in each category is orthogonal to each other and can be combined.

3.1.1. Noise Filtering. Noises are generally from in-band and out-of-band interference. Out-of-band interference is easy to be filtered out via digital filters such as Finite Impulse Response (FIR). In-band interference is usually hard to be removed. But SNR can still be improved by adopting a matched filter.

Finite Impulse Response (FIR) filters are widely used digital filters since they are inherently stable, have a linear phase, and are flexible in shaping their frequency responses. Therefore, FIR filter is easy to be implemented and usually achieve good performance. The filtering process is accomplished by a weighted sum of finite prior samples. A faster implementation uses the convolution between the inputs and the filter coefficients. Though FIR filters have many advantages, for resource-constrainted IoT devices, they are computation intensive. As a result, Cascade Integrator Comb (CIC) filters are developed. CIC filters achieve computation efficiency via decimation (i.e., downsampling). In addition, the frequency response of a CIC filter exhibits unique features. As depicted in Fig. 2 (b), significant losses appear at certain frequency bins,



No audible noise

(a) No waveform reshape with guard interval

(b) Waveform reshape with guard interval

Audible noise

Audible noise is greatly mitigated

(c) No waveform reshape without guard interval

(d) Waveform reshape without guard interval

Fig. 3: Audible artifacts due to speaker diaphragm inertia can be mitigated by waveform reshaping and the insertion of guard interval. The transmitted signal is in the range of $18 - 22$ kHz. Audible noises are indicated by the vertical lines below 15 kHz.

which can be used to suppress specific interference [Wang et al. 2016]. Another popular choice are matched filters. A matched filter can extract a known waveform in noise contaminated signals with low SNR by correlating the measurements with a known reference signal.

The above noise filtering techniques are performed at receiver ends. At the transmitter side, careful design can also mitigate noise such as inter-symbol-interference (ISI). For instance, an effective method to mitigate ISI is to insert Guard Interval (GI) between consecutive signal transmissions, keeping the channel silent for a while. Since acoustic echoes in the aerial channels are subject to 25 dB loss after 10 ms [Tung and Shin 2016], inserting GI can significantly reduce the impact of multipath reverberations from prior signals and thus mitigate ISI.

3.1.2. Channel Distortion Mitigation. The channels acoustic signals propagate through, are not ideal and often introduce distortions. There are two common source of channel distortions in acoustic systems, i.e., frequency selectivity [Mao et al. 2017] and speaker diaphragm inertial [Ine 1959].

Frequency selectivity, also known as non-flat frequency response, describes the phenomenon where acoustic signals experience different channel gains at different frequencies. It is common on commercial-off-the-shelf IoT devices since the acoustic sensors on these platforms are optimized only for the audible bandwidth [Lee et al. 2015;

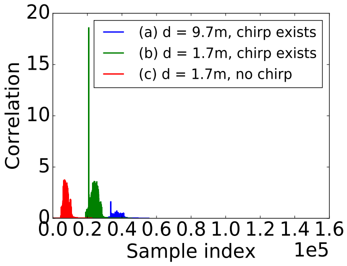


Fig. 4: The correlation peak of distant samples is even weaker than the noise peaks from close ones, making it challenging to set an appropriate threshold.

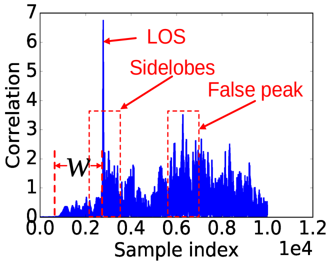


Fig. 5: Reliable onset detection illustration

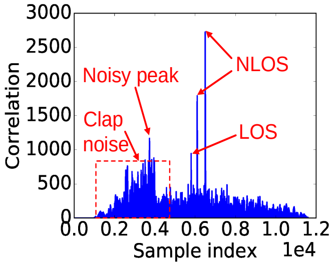


Fig. 6: Onset detection in challenging environment

Zhou et al. 2017] (normally below 8 kHz). However, signals at higher frequency bands are usually much favorable in acoustic sensing as they suffer less interference from background noises. As a result, a receiver will get corrupted inputs if they are not preprocessed. Frequency selectivity is often addressed by applying a compensation filter to the received signals with a reciprocal frequency response to the channels of interest.

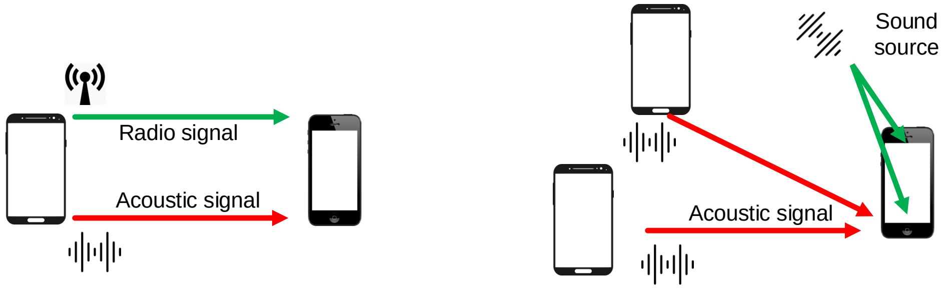
Speaker diaphragm inertia can cause ringing effects [Lee et al. 2015] or frequency leakage. Ringing effects are the distortions in time domain where a transmission is delayed at start and the duration of the transmission is prolonged. In contrast, frequency leakage (hereafter, we will use frequency leakage to refer to speaker diaphragm problem) describes the problem in frequency domain where a transmission of a band-limited signal can cause out-of-band noises. Perceptually, the speaker diagram inertial can generate audible noise though the transmitted signal only occupies inaudible frequency bands. Such a problem appears when a transmission has abrupt amplitude or phase changes. To address the problem, waveform reshaping techniques and channel estimation approaches have been considered in the literature. Waveform reshaping, as its name refers, mitigates channel distortions by slightly changing the waveform of the inputs. Existing solutions include utilizing a raised cosine window to reshape the waveforms [Lee et al. 2015], inserting fade-in and fade-out signals to ensure phase consistency [Lazik and Rowe 2012], or just slowly increasing and decreasing the amplitude of the first and last few samples [Zhou et al. 2017]. Fig. 3 demonstrates the effectiveness of waveform reshaping. Though waveform reshaping can reduce or eligible audible artifacts, it may introduce more distortions. Another technique, originated from RF communication systems, addresses the distortions by directly measuring channel responses [Roy et al. 2017]. After the channel responses are obtained, a reciprocal compensation filter can be designed and applied to the inputs. Since frequency leakage is more precisely compensated, distortions to the original signals are minimized. This, however, comes at the cost of higher implementation complexity for calibration.

3.1.3. Robust Onset Detection. Onset detection, determining the presence or absence of a particular signal and the associated timing, is the cornerstone of many acoustic sensing systems, in particular for time-sensitive applications.

Onset detection can be accomplished by naive FFT analysis or application of matched filters. FFT analysis is commonly used for tone detection. It achieves onset detection by inspecting if there is any known spectra appearing in the frequency domain while a matched filter accomplishes the task in time domain. Matched filtering is commonly applied to detect signals with good compression properties such as chirp signals. By cross-correlating the captured samples with a known reference signal, one can determine the presence and the timing of the reference signal from a strong peak. Anyway, both approaches require proper thresholds for peak detection. If the magnitude of the known spectra or the correlation peaks are above a certain threshold, the reference signal is identified, and vice versa. If multiple peaks are present, the maximum one is often chosen. Onset detection using peak detection with fixed thresholds is inadequate in dynamic and mobile situations due to the "near-far" effect, strong interference, and the multipath effect, which contribute to system failures.

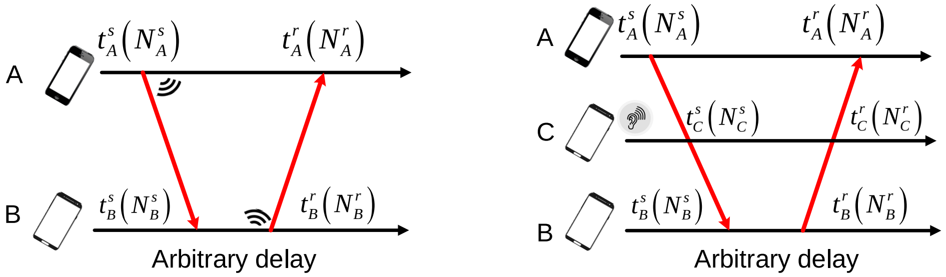
The "near-far" effect, a terminology originated from wireless communication systems, describes the phenomenon where the signal power received at a base station is dominated by closer users than the distant ones due to signal attenuation over distance. Acoustic sensing systems also suffer from the same problem. Specifically, the "near-far" problem makes it challenging to set an appropriate threshold to detect the reference signal at both near and far distances. When the threshold is high, distant signals may be missed while if the threshold is low, noise or interference near the receiver may be identified as the reference signals. Fig. 4 illustrates this dilemma. Moreover, when the received signals are saturated by strong interference with wide bandwidth (e.g., clap noise depicted in Fig. 6), the threshold-based detection method is problematic. A strong interference can easily generate multiple peaks exceeding the pre-set threshold, resulting in false onset detection. Finally, the multipath effect describes a phenomenon that a receiver not only captures the LOS signal that is assumed to be a predominant signal, but also receives multiple delayed and attenuated copies. These delayed and attenuated copies, called NLOS signals, can add up constructively so as to dominate the received signal. Consequently, in threshold-based onset detection systems, the timestamp corresponding to NLOS signals may be falsely identified as that of the LOS signal.

For robust detection, more sophisticated characteristics should be exploited. Fig. 5 illustrates some useful features for onset detection employing chirp signals. For instance, the ratio between the magnitude of an authentic peak and the mean value of its sidelobes [Peng et al. 2007] is much higher than that from interfering signals. Furthermore, when the reference signal appears, the ratio between the authentic peak and the mean of W samples ahead of the peak (depicted in Fig. 5) increases drastically while false peaks do not have this property. Utilizing the ratio (the latter one of the afore-mentioned two ratios) to normalize the cross-correlation results can effectively mitigate the "near-far" problem. However, in a multipath rich environment, multiple peaks, mainly generated by NLOS reflections, may also exceed a pre-defined threshold, making it hard to perform reliable onset detection. To mitigate this problem, one feasible approach [Peng et al. 2007] is to choose the peak that appears first since the LOS signal has a shorter propagation path than reflected signals. Another viable approach first computes the first order differences of the remaining peaks and then chooses the maximum one. In some cases, to extract desired signals from multiple reflections, application-specific features can be exploited. For instance, to retrieve the predominant echo from a moving finger, all the above methods will fail. In this specific scenario, finger-generated echoes exhibit dynamic features such as Doppler



(a) ToA via one way sensing (b) TDoA via one way sensing

Fig. 7: One way sensing



(a) ToA via two-way sensing (b) TDoA via two-way sensing

Fig. 8: Two way sensing

shift [Gupta et al. 2012] and phase changes [Wang et al. 2016; Yun et al. 2017], while other reflections do not have such properties.

3.2. Timing Estimation

Timing estimation aims to obtain the flight time of acoustic signals such as ToA or TDoA timestamps, which are critical especially for ranging and localization. ToA estimation usually involves two devices. It measures the absolute travel time of acoustic signals between transceiver pairs (Fig. 7 (a)). In contrast, TDoA typically involves multiple transceiver pairs and calculates the time difference (Fig. 8 (a)) instead. To perform ToA or TDoA estimation, existing solutions can be classified into one-way or two-way sensing approaches.

One-way sensing generally refers to the sensing paradigm where signal transmission has only one way from one or multiple transmitters to one or multiple receivers. One-way sensing often requires tight synchronization. For ToA estimation, this approach often exploits another high speed signal source (e.g., radio signals such as WiFi, Bluetooth, and Zigbee), with negligible travel time (compared to acoustic waves) for synchronization [Uddin and Nadeem 2013; Liu et al. 2013] as depicted in Fig. 7 (a). In this approach, a transmitter emits the acoustic and synchronization signals simultaneously. A receiver determines the ToA timestamp by computing the arrival time difference between the two signal sources. Therefore, a timestamp can be obtained without any coordination. For TDoA estimation, there are usually multiple transmitters or receivers. In some cases, the transmitters or receivers are physically on a single device. Either the transmitters or the receivers are tightly synchronized. In transmitter-synchronized systems, acoustic transmission are activated concurrently [Lazik and

Rowe 2012; Lazik et al. 2015]. TDoA is obtained by cross-correlating received samples with known reference signals. In receiver-synchronized systems, both actively generated acoustic waveforms and passive sounds [Liu et al. 2015; Zhu et al. 2014] (e.g., keystroke generated sound) can be used. TDoA is computed by cross-correlating the received samples from different channels. An illustration is given in Fig. 7. Note that mobile devices in the figures can be replaced by custom-built hardware, which allows more design flexibility and thus potentially achieves better performance. The main drawback of one-way sensing approaches lies in their needs for tight synchronization, which can be easily affected by system delay and network congestion.

Two-way sensing resolves timing information in a synchronization-free manner and thus is advantageous compared to one-way sensing. In two-way sensing, acoustic transmissions are bi-directional. Thus, a device needs to be equipped with both speaker and microphone. Fig. 8 (a) depicts the procedure to obtain ToA timestamps. At time t_A^s, device A starts an acoustic emission (usually a chirp signal). Device B detects the acoustic signal at time t_B^r and activates another transmission at time t_B^s after an arbitrary delay. Device A detects the second transmission at time t_A^r. Then ToA can be derived by the following equation [Peng et al. 2007]:

$$ToA_{AB} = \frac{1}{2}\left(t_A^r - t_A^s\right) - \frac{1}{2}\left(t_B^r - t_B^s\right). \tag{1}$$

If both transmissions can be received by a third device, then TDoA (depicted in Fig. 8 (b)) can be derived by [Wang et al. 2017]:

$$TDoA_{AB} = \frac{1}{2}\left(t_A^r - t_A^s\right) + \frac{1}{2}\left(t_B^r - t_B^s\right) - \left(t_C^r - t_C^s\right). \tag{2}$$

It is worth noting that all the timestamps, namely $t_A^s, t_A^r, t_B^r, t_B^s, t_C^r, t_C^s$, can be recorded as the sample indexes in an acoustic buffer, rather than the local system time, which is subject to various delays. Consequently, ToA or TDoA information can be efficiently obtained via sample counting. Furthermore, two-way sensing assumes the arbitrary delays depicted in Fig. 8 are the same across all the devices. This assumption is generally not true due to different sampling frequencies. Therefore, the arbitrary delay should be minimized. In some implementation, a coordinator is needed in two way sensing for scheduling the transmission, gathering all the timestamps for computation, and computing the final results for target devices.

Both one-way and two-way sensing are based on onset detection, which is often achieved via cross-correlation (CC). CC-based methods are subject to two- or three-samples error [Nandakumar et al. 2016], leading to an equivalent error of $1-2$ cm at a sampling rate $f_s = 48$ kHz and sound speed $c = 340$ m/s. Such a performance is insufficient for high-accuracy tracking such as finger tracking. As a result, two approaches have been devised in the literature for high accurate tracking.

The first method employs phase information [Nandakumar et al. 2016; Wang et al. 2016; Yun et al. 2017; Ke et al. 2018]. For instance, in a pure tone based system in which the tone signal oscillates at $f_c = 20$ kHz and the sampling rate is $f_s = 48$ kHz, a notable $\frac{\pi}{4}$ phase change is equivalent to fractional sample as $\frac{\frac{\pi}{4}}{2\pi} \times \frac{1}{f_c} = \frac{1}{8} \times \frac{1}{f_c} \approx \frac{1}{4} \times \frac{1}{f_s}$. Apparently, utilizing phase information enables finer timing resolution. Furthermore, it introduces less latency as accurate phase estimation can be done with only hundreds of samples [Wang et al. 2016]. High-precision tracking relies on finer-grained displacement estimation, which is computed as,

$$\frac{\theta}{2\pi} \times \frac{c}{f_c}, \tag{3}$$

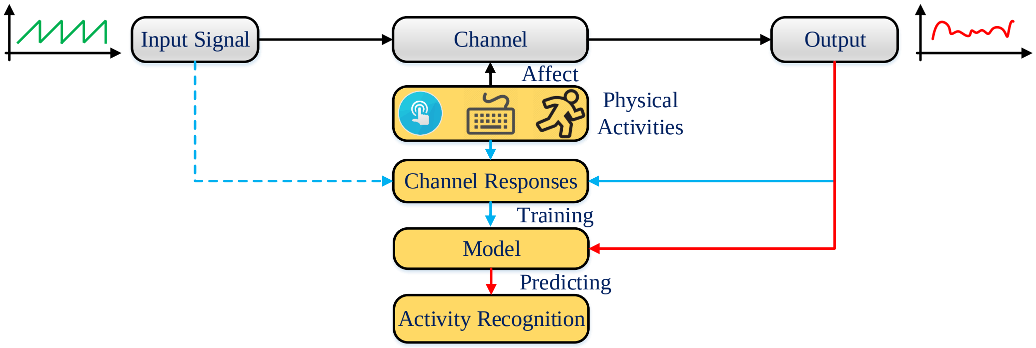


Fig. 9: Typical diagram for patten recognition

where θ is the accumulated phase shift and c is the sound speed. With more complex carriers like OFDM, more accurate estimation results can be obtained since they allow filtering outliers that single carrier systems are sensitive to. Other digital sequences such as GSM [Yun et al. 2017] and Zadoff-Chu sequences [Ke et al. 2018] can also infer displacement changes. Phase-based methods are more common in device-free gesture tracking systems.

The second method, used in device-based tracking, breaks the resolution barrier of CC by chirp signal mixing [Mao et al. 2016; Mao et al. 2017]. It works by translating displacement to frequency changes. The transceivers first perform one-time synchronization followed by tracking. Assuming the chirp at the receiver side is represented by $r = \cos\left(2\pi f_{\min} t + \pi \frac{B}{T} t^2\right)$, a displacement $d = c\Delta t$ results in signal delay Δt and attenuation $r_d = \alpha \cos\left(2\pi f_{\min}\left(t - \Delta t\right) + \pi k (t - \Delta t)^2\right)$, where f_{min} is the initial frequency, B is the bandwidth, T is the duration, α is the attenuation, and Δt is the elapsed time. Next, a signal mixing operation is performed by multiplying r and r_d. Taking the derivative of the mixed result with respect to t and filtering out high-frequency component, the delay or displacement is obtained as $\Delta t = \frac{fT}{B}$, where f is the frequency component of the remaining signal after mixing and filtering. f can be estimated by FFT analysis or advanced analytical models such as MUltiple SIgnal Classification (MUSIC). For example, with a bandwidth of $B = 4$ kHz, duration $T = 0.04$ s, and 1 Hz frequency estimation resolution, the equivalent resolution is $\frac{fT}{B} = 10^{-5} = 0.48 \times \frac{1}{f_s}$. Under appropriate settings, signal mixing can easily outperform CC-based methods in accuracy. Signal mixing operation has the additional benefit of robustness to the multipath effect. The multiple frequency components due to the multiple effect can be easily resolved by employing eigenvalue decomposition methods such as MUSIC and Singular Value Decomposition (SVD).

3.3. Pattern recognition

Pattern recognition aims to extract data regularities from raw measurements [Bishop 2006]. Acoustic sensing based on pattern recognition utilizes the fact that physical activities like keystroke, pressure, and breathing can either generate special acoustic signatures or affect the properties of acoustic channels. Some activities do not generate detectable acoustic signals themselves but affect acoustic channels and produce different channel responses. "channel" here refers to the medium that acoustic signals propagate through, which can either be the airspace or solid surfaces surrounding the sensing systems. Channel responses can be represented by a single scalar, i.e., signal strength, or high-dimensional features such as amplitude spectrum density and frequency spectrum.

Pattern recognition in active sensing systems usually involves three steps. Initially, acoustic signals spanning a wide bandwidth are transmitted through a channel influenced by different physical activities, and the channel responses are recorded. Through this process, sufficient labeled training data can be obtained. After that, a model is trained by mapping the channel responses to the target activities. Finally, the model is used online for pattern recognition based on the new samples. The canonical diagram of pattern recognition in active acoustic sensing systems is given in Fig. 9. In the first step, broadband signals such as chirps [Sur et al. 2014; Tung and Shin 2015] are commonly used since they can generate rich channel response signatures [Wang et al. 2014]. However, pure tone signals are also viable [Gupta et al. 2012]. In the second steps, statistical modeling or machine learning are common techniques to construct the model.

Statistical modeling associates the activities with quantifiable metrics, using a close-formed analytical model. For instance, in ForcePhone [Tung and Shin 2016], a vibrating phone is modeled as a forced and damped mass-spring system, where the relation between the applied force and the reduced vibration amplitude can be analytically represented. In SoundWave [Gupta et al. 2012], a closed-form inference model is applied to verify the presence of any Doppler frequency for gesture recognition. The analytical models are usually powerful and efficient but sensitive to measurement errors. Parameters in these models typically require calibration or training. Statistical modeling is non-trivial and usually requires domain knowledge. Consequently, machine learning models are gaining popularity.

Models employed in machine learning algorithms usually do not assume explicit functional relationships between target activities and acoustic-related features but assign different probabilities to activities associated with different acoustic profiles. Therefore, in the final predicting phase, the model often predicts the likelihood of different activities. Relevant machine learning methods include neural networks, decision trees, support vector machine (SVM), k-nearest neighbors (KNN), etc. Due to the time series nature of acoustic signals, they are first split into overlapping or non-overlapping segments. Then, time-domain or frequency-domain features are extracted from each segment as input to machine learning models in training or inference stages. Recently, there has been increasing application of deep models such as convolutional neural network models, recurrent neural network models in acoustic sensing systems to avoid the need for sophisticated hand-crafted features and achieve better classification performance.

Pattern recognition for passive sensing can also adopt the three-step procedure as that of active sensing. However, in passive sensing, acoustic signals are not purposefully generated but from natural sounds produced by corresponding physical activities. The intensity of these acoustic signals is usually very weak, and the patterns embedded in them are often buried under noise. Therefore, advanced signal processing techniques such as filtering [Nirjon et al. 2012] and signal transformation [Chauhan et al. 2017] are often employed before further processing.

3.4. Digital Modulation of Acoustic Signals

Digital modulation refers to the techniques that represent information as a function of carrier waves over a given medium [Proakis and Salehi 2008]. Conceptual, techniques in wireless communication systems such as source coding, channel coding, and modulation techniques as depicted in Fig. 10 can be applied. However, due to limited computation resources and available bandwidth, aerial acoustic communication tends to have much simpler design. For instance, sophisticated source coding and channel coding ap-

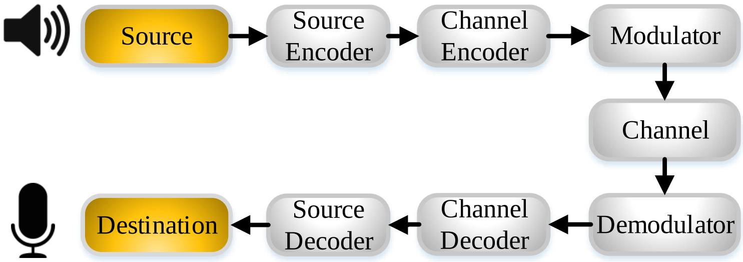


Fig. 10: Typical diagram of digital communication

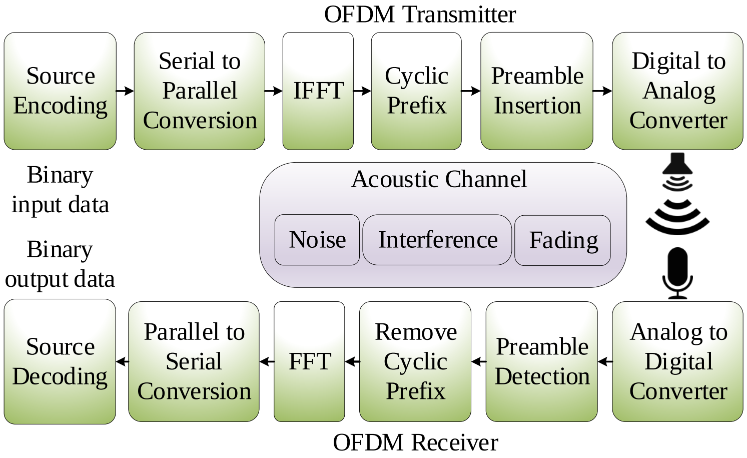


Fig. 11: Function block of baseband acoustic OFDM

proaches are seldom used. Channel coding methods with low complexity such as Cyclic Redundancy Check (CRC) [Wang et al. 2017], Forward Error Correction (FEC) [Wang et al. 2016] are more common. Considering the unique properties of acoustic signals (low propagating speed and low oscillating frequency), only a small portion of modulation techniques can be applied in acoustic sensing systems. Modulation techniques such as Phase Shift Keying (PSK) and Quadrature Amplitude Modulation (QAM) are seldom used as their performance degrades significantly from the Doppler shift [Lee et al. 2015]. As a result, in this section, we focus on the most commonly used modulation techniques, namely, FSK, OFDM, and chirp spread spectrum (CSS).

FSK is a modulation technique whereby pure tone signals with different frequencies are used to transmit data. Demodulating FSK signals can be done using FFT analysis, Hilbert Transform, or coherent detection. FSK is simple but can not achieve high throughput [Lopes and Aguiar 2001]. In contrast, OFDM is a more efficient modulation technique that deploys data symbols on orthogonal subcarriers and achieves high throughput with less bandwidth. However, since modulation is typically implementation in software in acoustic sensing platforms, acoustic OFDM structure is much simpler than its RF counter-part. Advanced signal processing modules in standard RF-based OFDM systems such as Carrier Frequency Offsets (CFO) correction, SFO correction, and carrier sensing have to be removed due to limited computation power. A simplified function block for acoustic OFDM is shown in Fig. 11 [Chapre et al. 2013]. First, bit streams are processed by channel coding techniques such as forward er-

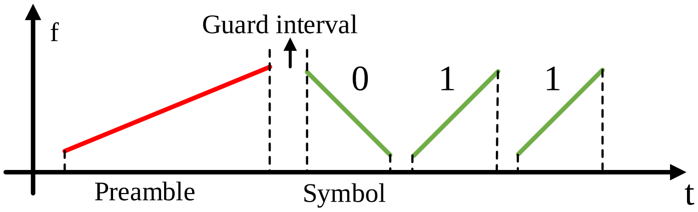


Fig. 12: An example of CSS frame

ror correction and cyclic redundancy check. This step adds redundant information to the original data streams and makes them more noise resilient. Afterwards, the bit streams are parallelized and go through Inverse Fast Fourier Transform (IFFT). This operation generates ready-to-transmit time domain signals. To mitigate inter-symbol-interference (ISI) and inter-channel interference (ICI), cyclic prefix/suffix (CP/CS) is inserted. CP/CS adds a repetition of generated signals. At this point, an OFDM frame is generated. To make the OFDM frame easy to be detected, a preamble (usually a chirp signal) is inserted before the packet. Finally, the signals are transmitted through the acoustic channels. A receiver reverses the above process.

Both FSK and OFDM are only suitable for short-range communication [Nandakumar et al. 2013] and suffer from degraded performance in mobile scenarios. In comparison, Chirp Spread Spectrum (CSS) [Kim and Chong 2015] is more reliable and can be used in long-range communication.

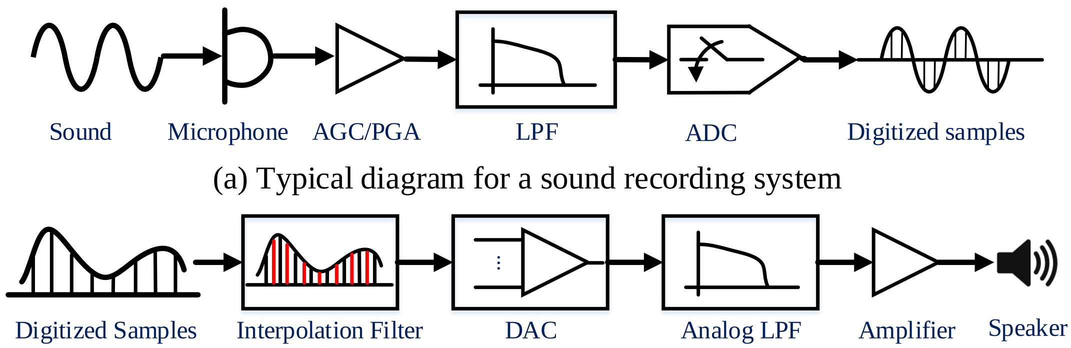
CSS is a known technique for LoRaWAN [LoRa Alliance 2015] (IEEE 802.15.4a) that aims to achieve long-range communication with low power consumption. It allocates wide bandwidth signals, namely chirp signals, to represent data symbols, making it robust to noisy interference and multipath fading. Also, it is suitable for mobile scenarios as chirp signals are resilient to the Doppler effect. A CSS frame often starts with a preamble followed by different data symbols. An example of a CSS frame is illustrated in Fig. 12. The preamble is used for synchronization and the data symbols are used to encode messages. Often, guard intervals are inserted between a preamble and data symbols to mitigate ISI. The CSS modulation has different symbol representation techniques. A good representation technique often leverages orthogonal chirps to denote different data symbols. Such a design can mitigate ISI and thus reduce the bit-error-rate. There are two well-known techniques, namely Binary Orthogonal Keying (BOK) [Lee et al. 2015] and Quaternary Orthogonal Keying (QOK) [Ka et al. 2016]. BOK utilizes orthogonal up- and down-chirps to represent different data symbols while QOK employs four orthogonal chirps. A CSS frame is decoded by matched filter at a receiving end. The main drawback of CSS modulation is its limited data rate.

4. PHYSICAL LAYER

Physical layer interfaces with acoustic hardware and the processing layer. It records audio streams and emits intended waveforms. In this section, we discuss physical layer design, in particular, major design issues including supporting hardware, coherent structure, waveform designs, and bandwidth.

4.1. Supporting hardware

Fig. 13 (a) and (b) depict the typical pipelines of sound recording [Roy et al. 2017] and emitting [Aud 2017] system, respectively. A sound recording system converts mechanical sound into digital samples while a sound emitting system reverses this process. In a recording system, sound signals are first converted into voltage signals by a mi-



(a) Typical diagram for a sound recording system

(b) Typical diagram for a sound emitting system

Fig. 13: Typical hardware diagram of acoustic sensing systems

crophone, the bandwidth of which is normally up to 100 kHz [Roy et al. 2017]. An Automatic Gain Control (AGC) or Programmable Gain Amplifier (PGA) then amplifies the voltage signals to surpass the quantization level of the posterior Analog-to-Digital Converter (ADC). Next, the amplified signals go through a Low Pass Filter (LPF), also known as an anti-aliasing filter, and become band-limited signals. The cut-off frequency of the LPF is $f_s/2$, where f_s is the sampling rate. The filtered signals go through a buffer and finally become digital samples via ADC conversion. A sound emitting system reverses the above process via a different circuit diagram. Digital samples are first interpolated, which are then fed into a Digital-to-Analog Converter (DAC) and become analog signals. The analog signals after amplification finally are converted to sound waves by a speaker. It should be noted that some platforms may adopt more than one microphone and speaker. For instance, modern smartphones utilize two speakers to play stereo audio and two microphones to enhance recording qualities.

4.2. Coherent detector

The recording diagram in Fig. 13 (a) can only obtain amplitude information. Extracting phase information needs more sophisticated design. A coherent detector depicted in Fig. 14 is a useful tool to extract the phase information [Wang et al. 2016; Yun et al. 2017]. In a coherent detector, an input signal creates two identical copies which are multiplied by $\cos{(2\pi ft)}$ and its $\frac{\pi}{2}$ phase shifted version $-\sin{(2\pi ft)}$, respectively. After a low pass filter, the In-phase (I) and Quadrature-phase components can then be obtained. Sometimes, interpolation and decimation [Wang et al. 2016; Yun et al. 2017] are applied, balancing the tradeoff between responsiveness and accuracy. It is worth noting that Fig. 14 only presents a simplified version of standard coherent detectors in RF field and omits many complex components [Tse and Viswanath 2005]. Off-the-shelf IoT devices, in absence of sophisticated hardware, usually adopt software defined coherent detector [Wang et al. 2016; Yun et al. 2017].

4.3. Waveform design

Waveform design is critical for acoustic sensing systems. A good waveform design can often bring better system performance. There are various acoustic waveforms. Among the available ones, chirp signals and pure tone signals are the most commonly used ones.

The chirp signals [Ka et al. 2016; Peng et al. 2007; Ruan et al. 2016; Zhou et al. 2017] (also known as Linear Frequency Modulated (LFM) signal), are represented by

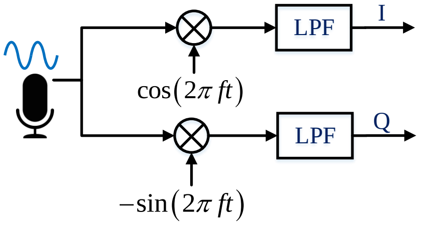


Fig. 14: Structure of coherent detector

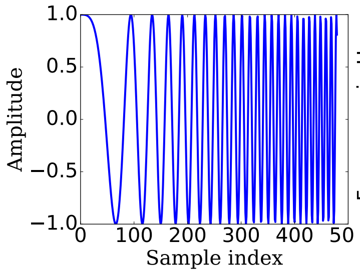


Fig. 15: Chirp signal in time domain

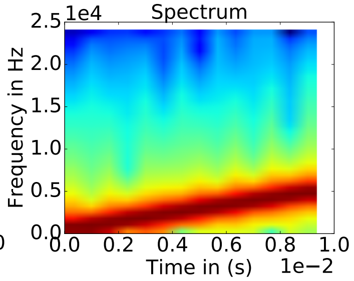


Fig. 16: Chirp signal in frequency domain

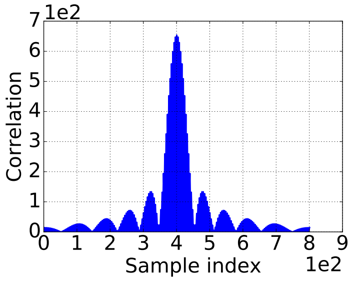


Fig. 17: Auto-correlation of chirp signal

$s\left(t\right)=A\cos\left(2\pi\left(f_{\min}t+\frac{k}{2}t^{2}\right)+\phi\right)$, where, f_{min} is the initial frequency, k is the modulation coefficient or chirp-rate, t is time, ϕ is the initial phase, and A is the maximum amplitude. The time and frequency domain representations of a chirp signal are depicted in Fig. 15 and Fig. 16, respectively.

Acoustic sensing systems employing chirp signals enjoy multi-fold benefits. First, chirp signals have good auto-correlation properties and are resilient to the Doppler effects. The good auto-correlation property, also known as Pulse Compression, can improve ranging resolution and receiver sensitivity. Such properties make chirp signals detectable even under noise floor [Ka et al. 2016]. Second, chirp signals is resilient to multipath fading. It is feasible to distinguish multiple reflections with appropriate signal processing models [Mao et al. 2017]. Third, it is easy to use chirp signals to design orthogonal signals by varying chirp-rates and bandwidth as in CSS modulation. The chirp rate k, i.e., modulation coefficient, can be positive, negative, or even vary with time. An example design of orthogonal chirps is shown in Fig. 18 (b). Chirp signals can also be pieced together, forming another widely used signals called Frequency Modulated Continuous Wave (FMCW, depicted in Fig. 18 (a)). Due to the aforementioned desirable properties, chirp signals have been widely used for synchronization, long-range communication, etc.

Pure tone signals or multiple carriers are another popular choices for acoustic sensing systems. A pure tone signal is denoted by $s\left(t\right)=cos\left(2\pi ft+\phi\right)$, where f is the frequency and ϕ is the phase. Pure tone signals also have many good properties. First, pure tone signals provide good resolutions to track Doppler shifts. Consider that a moving object transmits a pure tone signal at a frequency of f, and the detected Doppler frequency is $f_{shifted}$ at the receiver end. The moving speed can be estimated by $v=\frac{f_{shifted}-f}{f}c$, where c is the sound speed. Due to the low propagation speed of sound, it is feasible to achieve cm/s estimation accuracy. Second, pure tone signals enable fast and precise phase estimation. If multiple phase components across multiple

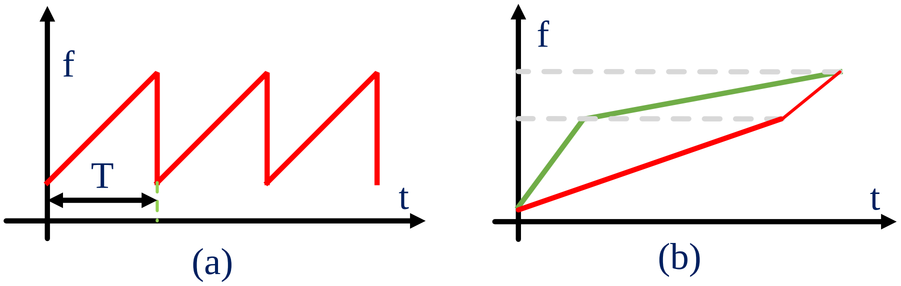


Fig. 18: Fig (a) depicts the waveform of the FMCW signal and Fig (b) depicts orthogonal chirp signal with different modulation coefficients

carriers are available, the phase divergence across each carrier can be formulated as a Chinese Remainder Theory problem [Lipson 1971; Goldreich et al. 1999], facilitating range or localization applications [Vasisht et al. 2016; Wang et al. 2016]. Finally, the frequency component of a pure tone signal can be easily detected and accurately estimated, enabling tone-based modulation. Owing to these advantages, pure tone signals have been extensively used in a wide range of applications such as tracking [Yun et al. 2015; Mao et al. 2016] and gesture recognition [Gupta et al. 2012; Wang et al. 2016; Ke et al. 2018].

Another types of viable waveforms are Zadoff-Chu (ZC) sequences [Stefania et al. 2011; Gul et al. 2015; Gul et al. 2012; Hyder and Mahata 2017]. A ZC sequence is the Primary Synchronization Signal (PSS) in LTE systems [Stefania et al. 2011], which can be represented by

$$seq\left(m+1\right)=e^{-j\pi Rm(m+1)/N}, m=0,...,N-1, \tag{4}$$

where R is the root and N is the sequence length. R and N are coprime integers. A ZC signal has many good properties. For instance, it has constant amplitudes and thus avoids the peak-to-average-power-ratio (PAPR) problem [Blcskei 2004; Heiskala and Terry 2001]. In addition, it is easy to obtain its frequency domain representation by efficient conjugate and scaling operations. One important characteristic is that a ZC signal is orthogonal to its delayed versions, which makes it perfect for synchronization. Furthermore, the real part of the ZC sequence also maintains this orthogonality. As depicted in Fig. 19, correlating a ZC signal with its noise-free version (a common approach for synchronization) results in only a single peak. In contrast, when using chirp signals, the correlation output has lots of comparable sidelobes as depicted in Fig. 17, which leads to synchronization errors [Nandakumar et al. 2016] especially in a time-varying channel. Apparently, utilizing ZC signals for synchronization outperforms chirp-based approaches. However, it should be noted that a ZC sequence, as indicated in Eq. 4, is phase-based and thus is not robust to Doppler effects. Doppler shifts in acoustic domains are more severe than that of RF domains, making it infeasible to use a ZC sequence in mobile scenarios.

4.4. Bandwidth consideration

Bandwidth is an essential design factor. The bandwidth of acoustic signals spans from 0.01 Hz to terahertz [White 1998]. However, commercial-off-the-shelf IoT devices usually limit the available bandwidth of acoustic transceivers to a frequency range of $0-24$ kHz or lower in order to prevent the anti-aliasing effect (one can extend the bandwidth and sampling rate by adopting custom-built hardwares). The bandwidth between $0-24$ Khz can be further broken down into *audible* and *inaudible* frequency

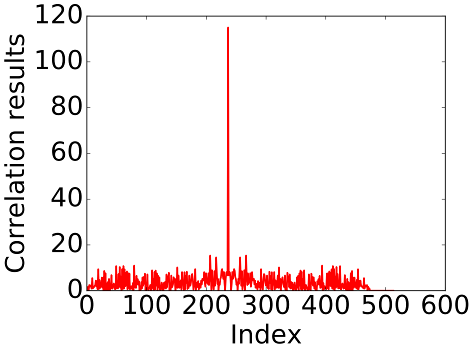


Fig. 19: Correlation results using a ZC signal

part. The audible part roughly spans the frequency range of $20 - 18000$ Hz, and the inaudible part occupies 6 kHz bandwidth, ranging from 18 to 24 kHz. The inaudible part, also known as near-ultrasound, is favorable for most acoustic sensing systems since it causes no disruptive audible noises and is less interfered by background noises. However, near-ultrasound signals are subject to more channel distortions due to frequency selectivity as discussed in Section 3.1. The channel response of acoustic sensors typically peaks around $6 - 7$ kHz [Wang et al. 2016] and decreases rapidly at higher frequency. Therefore, for better signal quality, signals band-limited to $6 - 7$ kHz are preferable.

The choice of bandwidth plays an important role in system performance. For instance, the bandwidth can affect the performance of timing estimation. To obtain accurate timing information, one needs to correctly determine the onset of a particular reference signal. Usually, this is done by correlating the captured signals with a reference signal and choosing the index of the maximum correlation peak as the onset as discussed in Section 3.1. Assuming the reference signal is the chirp signal represented by $s\left(t\right) = cos\left(2\pi f_{min} t + \pi \frac{B}{T} t^2\right)$, the correlation result is given by

$$\kappa\left(t\right) = T\left(1 - \frac{|t|}{T}\right) \sin c\left(\pi B\left(1 - \frac{|t|}{T}\right)\right) \cos\left(2\pi f_{\min} t\right). \tag{5}$$

Eq. 5 is in a form of cosine function modulated by a time-decaying sinc function [Cook 1974] shown in Fig. 17. From Eq. 5 we see that increasing the bandwidth B can suppress both the peak value and width of the sidelobes, thus reducing timing estimation errors. As another example, bandwidth can affect the resolution of timing estimation. As presented in Section 3.2, in high-accuracy tracking systems that adopt chirp mixing, the tracking resolution is defined by $\Delta t = \frac{fT}{B}$. Apparently, a larger bandwidth gives a finer resolution. However, with increased bandwidth, the problem of frequency selectivity becomes more severe and impairs system performance. As a result, tradeoffs must be made in judiciously selecting the bandwidth of employed acoustic signals.

5. CHALLENGES AND FUTURE DIRECTIONS

The research community so far has made much progress in the design of acoustic sensing systems. Many applications have been explored, some of which are becoming available in commercial products. However, there are still many challenging issues and untapped potentials in this field. In this section, we highlight the challenges and share our vision for future research trends.

5.1. Challenges

5.1.1. User configuration. End users typically prefer solutions that work directly out of box without any lengthy system setup, pre-training, or calibration. However, many existing solutions cannot satisfy these requirements. For instance, localization systems in [Lazik and Rowe 2012; Liu et al. 2013] require calibrating anchor nodes' coordinates; touch force [Tung and Shin 2016] enabling systems require pre-calibration, which may be difficult for non-professional end users. Therefore, eliminating or automating the setup process in practical deployment is needed.

5.1.2. Multipath effects. Multipath describes a phenomenon whereby acoustic signals reach a receiver through different propagation paths. It can hurt the performance of many acoustic sensing systems. For instance, the accuracy of range estimation depends on the presence detection of a dominating LOS path signal. However, in a multipath rich environment, the power of LOS signals may be much weaker than that of NLOS echoes since the latter can add up coherently or due to directionality of the transceivers. As a result, it is challenging to detect a LOS signal reliably. Though many multipath mitigation techniques have been proposed, they are often based on restrictive assumptions and are not robust to interference. For instance, in around device tracking systems, state-of-the-art work [Nandakumar et al. 2016; Wang et al. 2016; Yun et al. 2017] assumes that there is only a single dominating echo from the tracked object (e.g., a finger). Apparently, this assumption is not always true in practice. Consequently, more robust and precise multipath models are needed.

5.1.3. Sampling frequency offset. Due to the SFO between a transceiver pair [Kinoshita and Nakatani 2013; Miyabe et al. 2013a; Miyabe et al. 2013b], the duration of a specific signal can be different at two sides. Evidently, it is problematic especially for synchronization where timing precision is critical. The SFO problem is mainly caused by the instability of local oscillators [Miyabe et al. 2013b] which exhibit unpredictability and are susceptible to temperature changes [Miyabe et al. 2013b; Kinoshita and Nakatani 2013; Miyabe et al. 2013a]. Some existing solution assumes that SFO introduces a linear effect [Mao et al. 2016] and thus can be compensated. However, linear compensation is only applicable in a short period. Therefore, how to cope with the SFO problem in the long run still warrants investigation.

5.1.4. Heterogeneity. There are many acoustic sensor types with diverse sensitivity and frequency responses. Device diversity problem can hamper the scalability of solutions that rely on specific sensor properties. For instance, in BeepBeep [Peng et al. 2007], a threshold parameter, which is highly dependent upon the sensitivity of the microphones used, is needed to determine ToA timestamps. However, the heterogeneity of microphone sensitivity requires different thresholds for different sensing platforms. As a result, calibration is needed when running BeepBeep on different sensing platforms. Predictive models for pattern recognition, obtained through time-consuming training, are closely related to the frequency response of adopted acoustic sensors. Device heterogeneity makes a model trained using data from one device unsuitable for other devices. Device heterogeneity is often addressed through laborious device-dependent calibration, which is undesirable in real-world applications. Combining transfer learning with a few labeled measurements [Virmani and Shahzad 2017] can be a promising approach.

5.1.5. System Delay. System delay is a phenomenon where an acoustic sensor is not responsive to requests from application programs. For instance, it has been reported that the latency between issuing an audio playback command in the user space and the actual time of transmitting the desired acoustic signals on an Android operating system can be up to 10 ms [Technology 2018]. System delay is harmful to the responsiveness and accuracy of timing critical applications. For instance, excessive delay can affect the synchronization performance of ALPS [Lazik et al. 2015], leading to localization errors. In ARABIS, the update rate to obtain a location fix is limited to system delay. This problem is mainly caused by the uncertainty of code execution in the user and kernel space, which exhibits high variance and is highly correlated with system loads. Therefore, it is non-trivial to model and compensate for this uncertainty. Directly implementing the systems in the kernel space [Sur et al. 2014; Uddin and Nadeem 2013] can sidestep the problem. However, such an approach is often unscalable since it requires cumbersome device-dependent kernel modifications. Therefore, effective techniques handling system delays are needed, or the solutions should be made agnostic to such delays.

5.2. Future Direction

5.2.1. Acoustic mixer. A recent study on acoustic sensing found that a recording system can act as an acoustic mixer [Roy et al. 2017], making it possible to detect ultrasonic signals above 24 kHz on commercial-off-the-shelf mobile devices with no more than 48 kHz sampling rate. This phenomenon, caused by the non-linearity of acoustic sensors, has been exploited in jamming and communication [Roy et al. 2017]. This interesting findings may encourage more innovate applications in the near future. Meanwhile, such a technology also poses security threats to smart IoT devices with audio input functionality such as Google Home [CNET 2018b] and Amazon Echo [CNET 2018a]. It can synthesize audio signals unperceived by humans to manipulate smart IoT devices [Zhang et al. 2017; Roy et al. 2018] and thus open doors for malicious attacks. Therefore, techniques to detect and defense against such attacks are worthy of investigation.

5.2.2. Deep Learning. Recent years have witnessed a surge of deep learning [Goodfellow et al. 2016; LeCun et al. 2015]. Deep learning allows extracting useful features from end-to-end training. It even surpasses human performance in some taskes including image classification [Simonyan and Zisserman 2014; He et al. 2015], speech recognition [Taigman et al. 2014; Sun et al. 2014], etc. We believe that deep learning techniques can find many applications in acoustic sensing as well. For instance, convolutional network and recurrent neural network models can be used to recognize gestures from time series acoustic signals. Denoising AutoEncoders, which is originally used to learn the corrupted version of inputs, may be used to handle channel distortions and multipath effects.

6. CONCLUSION

In this paper, we presented a comprehensive survey on acoustic sensing. Based on the survey of existing work, we developed a layered architecture for acoustic sensing systems. This architecture encompasses three layers, namely, application layer, processing layer, and physical layer. In the application layer, we discussed three categories of enabled applications, including context-aware application, human-computer interface, and aerial acoustic communication. In the processing layer, different sensing approaches are analyzed comprehensively. In the physical layer, fundamental design considerations are presented in details.

Despite tremendous developments in acoustic sensing, there are still many technological challenges need further investigation, i.e., user configuration, multipath effect, sampling frequency offset, heterogeneity, and system delay. We believe that solutions to these challenges will not only improve system performance but also lead to a rise in many exciting applications. At the end of the survey, we introduced two hot research topics, namely acoustic mixer and deep learning. By the timely and thorough review of existing work, this survey may serve as guidelines and encourage more research efforts into acoustic sensing.

REFERENCES

1959. Loud speaker diaphragm. In *US Patent 2,905,260*.

2017. Apple iPhone6s 3DTouch. http://www.apple.com/iphone-6s/3d-touch/. (2017).

2017. TI Smartphone Solutions. http://www.ti.com/lit/sl/slyy032a/slyy032a.pdf. (2017).

H. Abdelnasser, M. Youssef, and K. A. Harras. 2015. WiGest: A ubiquitous WiFi-based gesture recognition system. In *Proceedings of the 34th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM) (INFOCOM '15)*.

N. Abhayasinghe and I. Murray. 2014. Human activity recognition using thigh angle derived from single thigh mounted IMU data. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN) (IPIN '2014)*.

Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C. Miller. 2014. 3d tracking via body radio reflections. In *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI) (NSDI '2014)*.

A. Berni and W. Gregg. 1973. On the Utility of Chirp Modulation for Digital Signaling. *IEEE Transactions on Communications* 21, 6 (Jun 1973), 748–751.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*.

S. S. Blackman. 1986. *Multiple-target tracking with radar applications*.

S. Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27, 2 (1979), 113–120.

Helmut Blcskei. 2004. Principles of MIMO-OFDM Wireless Systems. (2004).

Y. Chapre, P. Mohapatra, S. Jha, and A. Seneviratne. 2013. Received signal strength indicator and its analysis in a typical WLAN system (short paper). In *38th Annual IEEE Conference on Local Computer Networks*. 304–307.

Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing Acoustics-based User Authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

J. C. Chen, Kung Yao, and R. E. Hudson. 2002. Source localization and beamforming. *IEEE Signal Processing Magazine* 19, 2 (Mar 2002), 30–39.

Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. 2015. P-CNN: Pose-based CNN Features for Action Recognition. In *2015 IEEE Conference on Computer Vision (ICCV) (ICCV '15)*.

S. R. Cloude and E. Pottier. 1996. A review of target decomposition theorems in radar polarimetry. *IEEE Transactions on Geoscience and Remote Sensing* 34, 2 (Mar 1996), 498–518.

S. R. Cloude and E. Pottier. 1997. An entropy based classification scheme for land applications of polarimetric SAR. *IEEE Transactions on Geoscience and Remote Sensing* 35, 1 (Jan 1997), 68–78.

CNET. 2018a. Amazon Echo. https://www.cnet.com/products/amazon-echo-review/. (2018).

CNET. 2018b. Google Home. https://www.cnet.com/products/google-home/review/. (2018).

C. E. Cook. 1974. Linear FM Signal Formats for Beacon and Communication Systems. *IEEE Trans. Aerospace Electron. Systems* AES-10, 4 (July 1974), 471–478.

Scott Counts and Eric Fellheimer. 2004. Supporting Social Presence Through Lightweight Photo Sharing on and off the Desktop. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*.

S. E. El-Khamy, S. E. Shaaban, and E. A. Tabet. 1996. Efficient multiple-access communications using multi-user chirp modulation signals. In *Spread Spectrum Techniques and Applications Proceedings, 1996., IEEE 4th International Symposium on*, Vol. 3. 1209–1213.

B. Fernando, E. Gavves, M. Jos Oramas, A. Ghodrati, and T. Tuytelaars. 2015. Modeling video evolution for action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '15)*.

David Frohlich, Allan Kuchinsky, Celine Pering, Abbe Don, and Steven Ariss. 2002. Requirements for Photoware. In *In Proceedings of the ACM Computer Supported Cooperative Work (CSCW '02)*.

V. Gerasimov and W. Bender. 2000. Things that talk: Using sound for device-to-device and device-to-human communication. *IBM Systems Journal* 39, 3.4 (2000), 530–546.

Mayank Goel, Elliot Saba, Maia Stiber, Eric Whitmire, Josh Fromm, Eric C. Larson, Gaetano Borriello, and Shwetak N. Patel. 2016. SpiroCall: Measuring Lung Function over a Phone Call. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*.

Oded Goldreich, Dana Ron, and Madhu Sudan. 1999. Chinese Remaindering with Errors. In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing (STOC) (STOC '99)*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org

D. Graham, G. Simmons, D. T. Nguyen, and G. Zhou. 2015. A Software-Based Sonar Ranging Sensor for Smart Phones. *IEEE Internet of Things Journal* 2, 6 (2015), 479–489.

Malik Muhammad Usman Gul, Sungeun Lee, and Xiaoli Ma. 2012. Robust synchronization for OFDM employing Zadoff-Chu sequence. In *2012 46th Annual Conference on Information Sciences and Systems (CISS) (CISS)*.

M. M. U. Gul, X. Ma, and S. Lee. 2015. Timing and Frequency Synchronization for OFDM Downlink Transmissions Using Zadoff-Chu Sequences. *IEEE Transactions on Wireless Communications* 14, 3 (March 2015), 1716–1729.

Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. SoundWave: Using the Doppler Effect to Sense Gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*.

Michael Hanspach and Michael Goetz. 2014. On Covert Acoustical Mesh Networks in Air. *CoRR* abs/1406.1213 (2014). http://arxiv.org/abs/1406.1213

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*.

Juha Heiskala and John Terry, Ph.D. 2001. *OFDM Wireless LANs: A Theoretical and Practical Guide*.

Mashud Hyder and Kaushik Mahata. 2017. ZadoffChu Sequence Design for Random Access Initial Uplink Synchronization in LTE-Like Systems. *IEEE Transactions on Wireless Communications* 16, 1 (2017), 503–511.

Soonwon Ka, Tae Hyun Kim, Jae Yeol Ha, Sun Hong Lim, Su Cheol Shin, Jun Won Choi, Chulyoung Kwak, and Sunghyun Choi. 2016. Near-ultrasound Communication for TV's 2Nd Screen Services. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Sun Ke, Zhao Ting, Wang Wei, and Xie Lei. 2018. VSkin: Sensing Touch Gestures on Surfaces of Mobile Devices Using Acoustic Signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*.

Hamed Ketabdar, Kamer Ali Yüksel, and Mehran Roshandel. 2010. MagiTact: Interaction with Mobile Devices Based on Compass (Magnetic) Sensor. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI '10)*.

Sangdeok Kim and Jong-Wha Chong. 2015. Chirp Spread Spectrum Transceiver Design and Implementation for Real Time Locating System. *International Journal of Distributed Sensor Networks* 11, 8 (2015), 572861.

K. Kinoshita and T. Nakatani. 2013. Microphone-location dependent mask estimation for BSS using spatially distributed asynchronous microphones. In *2013 International Symposium on Intelligent Signal Processing and Communication Systems*. 326–331.

Heli Koskimäki, Pekka Siirtola, and Juha Röning. 2017. MyoGym: Introducing an Open Gym Data Set for Activity Recognition Collected Using Myo Armband. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp) (UbiComp '17)*.

Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: decimeter level localization using WiFi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM) (SigComm '2015)*.

Swarun Kumar, Stephanie Gil, Dina Katabi, and Daniela Rus. 2014. Accurate Indoor Localization with Zero Start-up Cost. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom) (MobiCom '2014)*.

Gierad Laput, Eric Brockmeyer, Moshe Mahler, Scott E. Hudson, and Chris Harrison. 2015. Acoustruments: Passive, Acoustically-driven, Interactive Controls for Handheld Devices. In *ACM SIGGRAPH 2015 Emerging Technologies (SIGGRAPH '15)*.

Eric C. Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N. Patel. 2012. SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*.

Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N. Patel. 2011. Accurate and Privacy Preserving Cough Sensing Using a Low-cost Microphone. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*.

Patrick Lazik, Niranjini Rajagopal, Oliver Shih, Bruno Sinopoli, and Anthony Rowe. 2015. ALPS: A Bluetooth and Ultrasound Platform for Mapping and Localization. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys '15)*.

Patrick Lazik and Anthony Rowe. 2012. Indoor Pseudo-ranging of Mobile Devices Using Ultrasonic Chirps. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys '12)*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. 521 (2015), 436–444.

H. Lee, T. H. Kim, J. W. Choi, and S. Choi. 2015. Chirp signal-based aerial acoustic communication for smart devices. In *2015 IEEE Conference on Computer Communications (INFOCOM '15)*.

Hong Li, Wei Yang, Jianxin Wang, Yang Xu, and Liusheng Huang. 2016. WiFinger: Talk to Your Smart Devices with Finger-grained Gesture. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) (UbiComp '16)*.

John D. Lipson. 1971. Chinese Remainder and Interpolation Algorithms. In *Proceedings of the Second ACM Symposium on Symbolic and Algebraic Manipulation (SYMSAC) (SYMSAC '71)*.

Cihang Liu, Lan Zhang, Zongqian Liu, Kebin Liu, Xiangyang Li, and Yunhao Liu. 2016. Lasagna: Towards Deep Hierarchical Understanding and Searching over Mobile Sensing Data. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Hongbo Liu, Yu Gan, Jie Yang, Simon Sidhom, Yan Wang, Yingying Chen, and Fan Ye. 2012. Push the Limit of WiFi Based Localization for Smartphones. In *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking (MobiCom '12)*.

Jian Liu, Yan Wang, Gorkem Kar, Yingying Chen, Jie Yang, and Marco Gruteser. 2015. Snooping Keystrokes with Mm-level Audio Ranging on a Single Phone. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*.

Kaikai Liu, Xinxin Liu, and Xiaolin Li. 2013. Guoguo: Enabling Fine-grained Indoor Localization via Smartphone. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '13)*.

C. V. Lopes and P. M. Q. Aguiar. 2001. Aerial acoustic communications. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*.

C. V. Lopes and P. M. Q. Aguiar. 2006. Acoustic communication system using mobile terminal microphone. 8, 2 (2006), 2–12.

Inc LoRa Alliance. 2015. LoRaWAN Specification. (2015).

Jonas Lowgren. 2014. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*

Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*.

Minghuang Ma, Haoqi Fan, and Kris M. Kitani. 2016. Going Deeper into First-Person Activity Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '16)*.

Wenguang Mao, Jian He, Huihuang Zheng, Zaiwei Zhang, and Lili Qiu. 2016. High-precision Acoustic Motion Tracking: Demo. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor Follow Me Drone. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

S. Mitra and T. Acharya. 2007. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 3 (May 2007), 311–324.

S. Miyabe, N. Ono, and S. Makino. 2013a. Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 674–678. DOI:http://dx.doi.org/10.1109/ICASSP.2013.6637733

S. Miyabe, N. Ono, and S. Makino. 2013b. Optimizing frame analysis with non-integrer shift for sampling mismatch compensation of long recording. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 1–4.

Chaithanya Kumar Mummadi, Frederic Philips Peter Leo, Keshav Deep Verma, Shivaji Kasireddy, Philipp Marcel Scholl, and Kristof Van Laerhoven. 2017. Real-time Embedded Recognition of Sign Language Alphabet Fingerspelling in an IMU-Based Glove. In *Proceedings of the 4th International Workshop on Sensor-based Activity Recognition and Interaction (iWOAR) (iWOAR '17)*.

Samuel Murray. 2017. Real-Time Multiple Object Tracking - A Study on the Importance of Speed. *CoRR* abs/1709.03572 (2017). http://arxiv.org/abs/1709.03572

T. Nadeem and L. Ji. 2007. Location-Aware IEEE 802.11 for Spatial Reuse Enhancement. *IEEE Transactions on Mobile Computing* 6, 10 (Oct 2007), 1171–1184.

Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, Venkat Padmanabhan, and Ramarathnam Venkatesan. 2013. Dhwani: Secure Peer-to-peer Acoustic NFC. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM (SIGCOMM '13)*.

Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N. Padmanabhan. 2012. Centaur: Locating Devices in an Office Environment. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*.

Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless Sleep Apnea Detection on Smartphones. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '15)*.

Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*.

Shahriar Nirjon, Robert F. Dickerson, Philip Asare, Qiang Li, Dezhi Hong, John A. Stankovic, Pan Hu, Guobin Shen, and Xiaofan Jiang. 2013. Auditeur: A Mobile-cloud Service Platform for Acoustic Event Detection on Smartphones. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '13)*.

Shahriar Nirjon, Robert F. Dickerson, Qiang Li, Philip Asare, John A. Stankovic, Dezhi Hong, Ben Zhang, Xiaofan Jiang, Guobin Shen, and Feng Zhao. 2012. MusicalHeart: A Hearty Way of Listening to Music. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys '12)*.

Makoto Ono, Buntarou Shizuki, and Jiro Tanaka. 2015. Sensing Touch Force Using Active Acoustic Sensing. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '15)*.

Esben Warming Pedersen and Kasper Hornbæk. 2014. Expressive Touch: Studying Tapping Force on Tabletops. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*.

Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th International Conference on Embedded Networked Sensor Systems (SenSys '07)*.

Lacomme Philippe, Hardange Jean-Philippe, and Jean-Claude Marchain. 2001. *Air and Spaceborne Radar Systems: An Introduction*.

H. Pirsiavash and D. Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '12)*.

Y. Prathivadi, J. Wu, T. R. Bennett, and R. Jafari. 2014. Robust activity recognition using wearable IMU sensors. In *IEEE SENSORS 2014 Proceedings*.

John G. Proakis and Masoud Salehi. 2008. *Digital Communications. 5th ed.*

Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home Gesture Recognition Using Wireless Signals. In *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking (MobiCom) (MobiCom '13)*.

Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*.

B. Rafaely. 2005. Analysis and design of spherical microphone arrays. *IEEE Transactions on Speech and Audio Processing* 13, 1 (Jan 2005), 135–143.

Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A Mobile System for Sensing Non-speech Body Sounds. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*.

Y. Ren, C. Wang, J. Yang, and Y. Chen. 2015. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In *2015 IEEE Conference on Computer Communications (INFOCOM '15)*.

Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI) (NSDI '18)*.

Wenjie Ruan, Quan Z. Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: Enabling Fine-grained Hand Gesture Detection by Decoding Echo Signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, 1 (1978), 43–49.

Annie Shoup, Tanya Talkar, Jian-Hua Chen, and Anubhav Jain. 2016. An Overview and Analysis of Voice Authentication Methods.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*.

Sesia Stefania, Toufik Issam, and Baker Matthew. 2011. *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*.

Ke Sun, Wei Wang, Alex X. Liu, and Haipeng Dai. 2018. Depth Aware Finger Tapping on Virtual Displays. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*.

Y. Sun, X. Wang, and X. Tang. 2014. Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*.

Sanjib Sur, Teng Wei, and Xinyu Zhang. 2014. Autodirective Audio Capturing Through a Synchronized Smartphone Array. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*.

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*.

Superpowered Audio Technology. 2018. Android Audio's 10 Millisecond Problem: The Android Audio Path Latency Explainer. http://superpowered.com/. (2018).

David Tse and Pramod Viswanath. 2005. *Fundamentals of Wireless Communication*. Cambridge University Press.

Yu-Chih Tung and Kang G. Shin. 2015. EchoTag: Accurate Infrastructure-Free Indoor Location Tagging with Smartphones. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*.

Yu-Chih Tung and Kang G. Shin. 2016. Expansion of Human-Phone Interface By Sensing Structure-Borne Sound Propagation. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '16)*.

M. Uddin and T. Nadeem. 2013. RF-Beep: A light ranging scheme for smart devices. In *2013 IEEE International Conference on Pervasive Computing and Communications (PerCom '13)*.

Deepak Vasisht, Swarun Kumar, and Dina Katabi. 2016. Decimeter-level Localization with a Single WiFi Access Point. In *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation (NSDI) (NSDI '2016)*.

Aditya Virmani and Muhammad Shahzad. 2017. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys) (MobiSys '17)*.

C. Wang, Y. Wang, and A. L. Yuille. 2013. An Approach to Pose-Based Action Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '13)*.

Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. 2014. Ubiquitous Keyboard for Small Mobile Devices: Harnessing Multipath Fading for Fine-grained Keystroke Localization. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*.

Qian Wang, Kui Ren, Man Zhou, Tao Lei, Dimitrios Koutsonikolas, and Lu Su. 2016. Messages Behind the Sound: Real-time Hidden Acoustic Signal Capture with Smartphones. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and Modeling of WiFi Signal Based Human Activity Recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom) (MobiCom '15)*.

Wei Wang, Alex X. Liu, and Ke Sun. 2016. Device-free Gesture Tracking Using Acoustic Signals. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Yu-Ting Wang, Jun Li, Rong Zheng, and Dongmei Zhao. 2017. ARABIS: an Asynchronous Acoustic Indoor Positioning System for Mobile Devices. In *2017 IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN 2017)*.

Y. Wang, R. Zheng, and D. Zhao. 2016. Towards Zero-Configuration Indoor Localization Using Asynchronous Acoustic Beacons. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*.

Shi-Yao Wei, Chen-Yu Wang, Ting-Wei Chiu, Yi-Ping Lo, Zhi-Wei Yang, Hsing-Man Wang, and Yi-ping Hung. 2016. RunPlay: Action Recognition Using Wearable Device Apply on Parkour Game. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST) (UIST '16)*.

R. M. White. 1998. Acoustic sensors for physical, chemical and biochemical applications. In *Proceedings of the 1998 IEEE International Frequency Control Symposium (Cat. No.98CH36165)*.

Y. Xiang, A. Alahi, and S. Savarese. 2015. Learning to Track: Online Multi-object Tracking by Decision Making. In *2015 IEEE International Conference on Computer Vision (ICCV)*.

Jie Xiong and Kyle Jamieson. 2013. ArrayTrack: A Fine-grained Indoor Location System. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation (NSDI '13)*.

Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time Tracking of Mobile RFID Tags to High Precision Using COTS Devices. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom) (MobiCom '2014)*.

Yanbing Yang, Jiangtian Nie, and Jun Luo. 2017a. ReflexCode: Coding with Superposed Reflection Light for LED-Camera Communication. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom '17)*.

Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao. 2017b. A Survey on Security and Privacy Issues in Internet-of-Things. *IEEE Internet of Things Journal* 4, 5 (Oct 2017), 1250–1258.

Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object Tracking: A Survey. *ACM Comput. Surv.* 38, 4 (Dec. 2006).

H. S. Yun, K. Cho, and N. S. Kim. 2010. Acoustic Data Transmission Based on Modulated Complex Lapped Transform. *IEEE Signal Processing Letters* 17, 1 (Jan 2010), 67–70.

Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a Mobile Device into a Mouse in the Air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '15)*.

Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS) (CCS '17)*.

Huanle Zhang, Wan Du, Pengfei Zhou, Mo Li, and Prasant Mohapatra. 2016. DopEnc: Acoustic-based Encounter Profiling Using Smartphones. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*.

Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. 2012. SwordFight: Enabling a New Class of Phone-to-phone Action Games on Commodity Phones. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys) (MobiSys '12)*.

Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic Sensing Based Indoor Floor Plan Construction Using Smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

Tong Zhu, Qiang Ma, Shanfeng Zhang, and Yunhao Liu. 2014. Context-free Attacks Using Keyboard Acoustic Emanations. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*.

E. ztrk, D. Genschow, U. Yodprasit, B. Yilmaz, D. Kissinger, W. Debski, and W. Winkler. 2017. A 60-GHz SiGe BiCMOS Monostatic Transceiver for FMCW Radar Applications. *IEEE Transactions on Microwave Theory and Techniques* 65, 12 (2017), 5309–5323.