SCANNING THE LITERATURE

The Scanning the Literature column provides concise summaries of selected papers that have recently been published in the field of networking. Each summary describes the paper's main idea, methodology, and technical contributions. The purpose of the column is to bring the state of the art of networking research to readers of *IEEE Network*. Authors are also welcome to recommend their recently published work to the column, and papers with novel ideas, solid work, and significant contributions to the field are especially appreciated. Authors wishing to have their papers presented in the column should contact the Editor.

Xiaohua Tian, Shanghai Jiao Tong University xtian@sjtu.edu.cn

Acoustic devices in various forms are becoming ubiquitous in Internet of Things (IoT) scenarios. The use of smart voice assistants and wearable devices such as Bluetooth headsets and smart microphones makes human-computer interfaces more user-friendly. Recent studies show that acoustic communication can go beyond mere interfacing functionality, where new application scenarios and technical challenges are revealed. The column in this issue focuses on recent research on acoustic communication in IoT. We pay special attention to topics including voice localization, defense of acoustic attacks, personalized headphones, and enabling spatial information in acoustic devices.

The popularity of smart speakers has grown exponentially over the past few years. The ability to localize human voice benefits smart speakers in many ways. There have been a number of interesting works on motion tracking and localization using audio, RF, and vision-based schemes. However, the existing schemes proposed require transmission of known signals and estimation of indoor environment, and the user has to be in line of sight (LoS) from the smart speaker. To address those challenges, Wang et al. present a novel system, MAVL, that can localize human voice in the following paper.

MAVL: Multiresolution Analysis of Voice Localization

Mei Wang, Wei Sun, and Lili Qiu, "MAVL: Multiresolution Analysis of Voice Localization," *Proc. 18th USENIX Symp. Networked Systems Design and Implementation*, 2021.

The authors first develop a novel multi-resolution analysis to estimate the angle of arrival (AoA) of time-varying low-frequency coherent voice signals coming from multiple propagation paths. The room structure is then estimated by emitting acoustic signals with a proposed 3D MUSIC algorithm. They finally re-trace the paths using the estimated AoA and room structure to localize the voice. The authors implement a prototype system using a single speaker and a uniform circular microphone array. Experimental results show that the proposed system achieves median errors of 1.49° and 3.33° for the top two AoAs' estimation and achieves median localization errors of 0.31 m in LoS cases and 0.47 m in non-LoS (NLoS) cases.

More than 3.25 billion voice assistants (e.g., Siri, Alexa) have been installed around the world, and it is anticipated that by 2023 the number will reach up to 8 billion. Researchers have identified various attacks against such systems, and one of the most devastating attacks is DolphinAttack, whereby attackers can inject inaudible voice commands and perform various malicious attacks, such as open a door, make a phone call, or place an order. To defend against DolphinAttacks, researchers have proposed several strategies, which, however, cannot defend against a sophisticated attacker and may induce health issues. To further study this issue, Zhang et al. propose a lightweight detection method, EarArray, which looks into the propagation difference between inaudible voice commands (i.e., ultrasound) and audible ones in the following paper.

EarArray: Defending against DolphinAttack via Acoustic Attenuation

Guoming Zhang, Xiaoyu Ji, Xinfeng Li, Gang Qu, and Wenyuan Xu, "EarArray: Defending against DolphinAttack via Acoustic Attenuation," *Proc. Network and Distributed Systems Security Symp.*, 2021.

DolphinAttacks (i.e., inaudible voice commands) modulate audible voices over ultrasound to inject malicious commands silently into voice assistants and manipulate controlled systems (e.g., doors or smart speakers). Eliminating DolphinAttacks is challenging if even possible since it requires modifying the microphone hardware. In this paper, the authors design EarArray, a lightweight method that can not only detect such attacks but also identify the direction of attackers without requiring any extra hardware or hardware modification. Essentially, inaudible voice commands are modulated on ultrasound that inherently attenuates faster than audible sounds. By inspecting the command sound signals via the built-in multiple microphones on smart devices, EarArray is able to estimate the attenuation rate and thus detect the attacks. They propose a model of the propagation of audible sounds and ultrasound from the sound source to a voice assistant (e.g., a smart speaker), and illustrate the underlying principle and its feasibility. They implemented EarArray using two specially designed microphone arrays, and their experiments show that EarArray can detect inaudible voice commands with an accuracy of 99 percent and recognize the direction of the attackers with an accuracy of 97.89 percent.

Humans can inherently sense the direction θ from which a sound arrives at their ears. The human brain essentially analyzes the time difference of the sounds across the two ears and maps this difference. Briefly, sounds that actually enter the ear canal are influenced by the shape of the human head and the pinna of the ear. One can view the head (including the pinna) as a filter that alters the signal depending on its angle of arrival θ . In the frequency domain, this filter is called head related transfer function (HRTF). Knowing HRTF for each θ opens new possibilities in spatial acoustics. However, estimating personalized HRTFs for each user is not achievable. To address the challenge, Yang et al. develop a personalization system, UNIQ, which combines techniques from channel estimation, motion tracking, and signal processing to estimate a user's personal HRTF in the following paper:

Personalizing Head Related Transfer Functions for Earables

Zhijian Yang and Romit Roy Choudhury, *Proc. Annual Con. ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2021.

Head related transfer functions (HRTFs) describe how sound signals bounce, scatter, and diffract when they arrive at the head, and travel toward ear canals. HRTFs produce distinct sound patterns that ultimately help the brain infer the spatial properties of the sound, such as its direction of arrival, θ . If an earphone can learn the HRTF, it could apply the HRTF to any sound and make that sound appear directional to the user. For instance, a directional voice guide could help a tourist navigate

SCANNING THE LITERATURE

a new city. While past works have estimated human HRTFs, an important gap lies in personalization. Today's HRTFs are global templates that are used in all products; since human HRTFs are unique, a global HRTF only offers a coarse-grained experience. This paper shows that by moving a smartphone around the head, combined with mobile acoustic communications between the phone and the earbuds, it is possible to estimate a user's personal HRTF. Their personalization system, UNIQ, combines techniques from channel estimation, motion tracking, and signal processing, with a focus on modeling signal diffraction on the curvature of the face. The results are promising and could open new doors into the rapidly growing space of immersive augmented/virtual reality (AR/VR), earables, smart hearing aids, and so on.

Advanced audio sensing and acoustic wearable technologies have been long awaited by mobile consumers, particularly spatial sound field technology, which offers a high-dimension acoustic experience with a broad spectrum of applications, such as 3D sound in AR/VR and sound source localization. Streaming service platforms are starting to support spatial sound experience as well. Spatial sound recording typically employs a microphone array to simultaneously record multitrack audio signals in different directions in order to reconstruct the sound field in a space. It is known that only with highly precise synchronization will the time-domain waveform differences between the microphone-array elements be accurately derived, which is the key information to calculate the sound source position and strength distribution. There have been significant related studies on sound synchronization in the past, but the evolution toward consumer wearable devices that are of small form factor and low power is full of challenges. To address those practical issues, Zhao et al. present a novel microphone array sensor architecture that enables synchronous concurrent transmission of multi-track audio signals using analog backscatter communication in the following paper.

Microphone Array Backscatter: An Application-Driven Design for Lightweight Spatial Sound Recording over the Air

Jia Zhao, Wei Gong, and Jiangchuan Liu, "Microphone Array Backscatter: An Application-Driven Design for Lightweight Spatial Sound Recording over the Air," *Proc. 28th Annual Int'l. Conf. Mobile Computing and Networking*, 2022.

Modern acoustic wearables with microphone arrays are promising to offer rich experience (e.g., 360° sound and acoustic imaging) to consumers. However, real-time multi-track audio streaming with precise synchronization poses significant challenges to the existing wireless microphone array designs that depend on complex digital synchronization as well as bulky and power-hungry hardware. This paper presents a novel microphone array sensor architecture that enables synchronous concurrent transmission of multitrack audio signals using analog backscatter communication. The authors develop novel pulse position modulation (PPM) and differential pulse position modulation (DPPM) baseband circuits that can generate a spectral-ef-

ficient, time-multiplexing, and multi-track-synchronous baseband signal for backscattering. Its lightweight analog synchronization supports parallel multimedia signals without using any ADCs, DSPs, codecs, or RF transceivers, hence largely reducing complexity, latency, and power consumption. To further enhance self-sustainability, they also design an energy harvester that can extract energy from both sound and RF. They have built a microphone array backscatter sensor prototype using an FPGA, discrete components, and analog devices. Their experiments demonstrate a communication range (sensor-to-reader) of up to 28 m for 8 audio tracks, and an equivalent throughput of up to 6.4 Mb/s with a sample rate over 48 kHz. Their sensor achieves 87.4µs of streaming latency for 4 tracks, which is 650× improvement compared to digital solutions. ASIC design results show that it consumes as low as 175.2 µW of power. Three sample applications, including an acoustic imaging system, a beamform filter, and a voice control system, all built with their phased-array microphone, further demonstrate the applicability of this design.

VR and AR technologies are developing rapidly and benefit various domains. There are two critical issues in VR and AR systems: to ensure that virtual objects appear in the correct places in the scene and another ensuring that the users perceive virtual objects accurately relative to other virtual and physical objects in a scene. To address the challenge, Garg et al. develop a virtual shuffleboard experimental scene to compare users' depth perception in VR and AR in the following paper.

Owlet: Enabling Spatial Information in Ubiquitous Acoustic Devices

Nakul Garg, Yang Bai, and Nirupam Roy, *Proceedings of the 19th International Conference on Mobile Systems, Applications, and Services*, 2021.

This paper presents a low-power and miniaturized design for acoustic direction-of-arrival (DoA) estimation and source localization called Owlet. The required aperture, power consumption, and hardware complexity of the traditional array-based spatial sensing techniques make them unsuitable for small and power-constrained IoT devices. Aiming to overcome these fundamental limitations, Owlet explores acoustic microstructures for extracting spatial information. It uses a carefully designed 3D-printed metamaterial structure that covers the microphone. The structure embeds a direction-specific signature in the recorded sounds. The Owlet system learns the directional signatures through one-time in-lab calibration. The system uses an additional microphone as a reference channel and develops techniques that eliminate environmental variation, making the design robust to noises and multipaths in arbitrary locations of operations. The Owlet prototype shows 3.6° median error in DoA estimation and 10° median error in source localization while using a $1.5 \text{ cm} \times 1.3 \text{ cm}$ acoustic structure for sensing. The prototype consumes less than 100th of the energy required by a traditional microphone array to achieve similar DoA estimation accuracy. Owlet opens up possibilities of low-power sensing through 3D-printed passive structures.

IEEE Network - January/February 2022