ABSTRACT

Title of Dissertation: SOUND SENSING, ENHANCEMENT, AND SEPARATION WITH MILLIMETER WAVE RADIO

Muhammed Zahid Ozturk
Doctor of Philosophy, 2022

Dissertation Directed by: Professor K. J. Ray Liu
Department of Electrical & Computer Engineering

Sound, as one the most natural way of human communication, has become a ubiquitous modality for human-machine-environment interactions. Despite many environmental sensing capabilities enabled by microphones, sound sensing systems have limitations, such as weak source separation when multiple speakers are present, being prone to replay attacks, and reduced performance under interference and noise. On the other hand, thanks to the availability of next generation communication systems and miniaturized radars, mmWave has become an emerging sensing modality in the recent years. Mobile phones and smart hubs include mmWave radars for environment sensing. To extend the sensing capabilities of these devices, and overcome limitations of microphones, we explore sound sensing and its applications by mmWave radars.

In this dissertation, we first explore how and to what extent ambient sound and sound-induced vibration could be sensed by mmWave-based sensing. We first establish fundamentals to sense sound from ambient objects, such as a piece of aluminum foil, or active speaker surfaces.

We show that, unlike microphones, which sense the sound at the sensor location, radars can sense sound remotely (*e.g. from the environment*), and robustly. We conduct a variety of experiments to understand the limitations of sound sensing from passive objects.

After establishing the fundamentals of sound sensing from the environment, we propose *RadioMic*, a system that can detect and localize the source of a sound robustly, and enhance the noisy radar signals via deep learning methods. Extensive experiments show how our proposal outperforms existing work and enables sound sensing in challenging conditions, such as through-wall and through-soundproof objects. Furthermore, *RadioMic* can extract individual sound streams when multiple sources are present. Last, we illustrate how *RadioMic* can detect whether a source is a live source or an inanimate source, mitigating the vulnerability of microphones against replay attacks.

Next, we investigate another limitation of microphone-based sensing, being prone to interference and noise. In other words, microphones usually have weak source separation capabilities, and recent deep learning based approaches do not perform well under challenging conditions. Furthermore, monaural speech separation has additional limitations, such as the problem of source association, and the number of speaker estimation. We build a system *RadioSES* that uses complementary radio modality to mitigate these fundamental drawbacks of microphones in speech enhancement and separation. Our extensive experiments indicate that *RadioSES* solves source association and tracking problems robustly, and improves the performance in speech enhancement and separation by 3 to 6 dB SiSDR, compared to the audio-only baseline. Furthermore, *RadioSES* can work in dark and through occlusion cases, and is preferable over using video modality, as it is less privacy concerning and computationally more efficient.

Last, we study the voice activity detection (VAD) problem using radio modality. Voice

activity detection is an integral part of smart speakers and voice transmission systems. A high-performance and automated VAD is of utmost importance, especially when the user intervention is limited, such as while driving a car. When the application requires focusing on *a particular user*, existing audio-based methods perform poorly as the interfering speakers or severe background noise create false alarms. We present *RadioVAD*, a radio-based VAD system that is robust against interference and noise. Our careful evaluation indicates that *RadioVAD* can match the performance of audio-VAD, at a much lower computational complexity, and can outperform existing approaches. Furthermore, we present different case studies to better understand the tradeoff between audio and radio SNRs, and investigate the false alarm, precision, recall rates, and detection delay carefully.

SOUND SENSING, ENHANCEMENT, AND SEPARATION WITH MILLIMETER WAVE RADIO

by

Muhammed Zahid Öztürk

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor K. J. Ray Liu, Chair/Advisor
Professor Gang Qu
Professor Min Wu
Dr. Beibei Wang
Professor Lawrence C. Washington (Dean's Representative)

Dedication

To my family, and my *spring*

Acknowledgments

 [1]

All praise belongs to Allah ﷻ, the Lord of the universes. Peace be upon the prophet of Allah ﷻ, Muhammed ﷺ, who is the best example and guide for all humanity.

First, I would like to thank my advisor, Prof. K. J. Ray Liu for his invaluable guidance, patience, motivation, and knowledge. His continuous support and encouragement in exploring unknown domains have empowered me to find cutting-edge research ideas, and execute those ideas successfully. I will always appreciate his advice and suggestions not only in research but also in life.

I would like to thank all of my dissertation committee members. I am grateful to Prof. Min Wu for her rigorous approach to new research ideas, and for bringing new aspects to research problems we encountered. I would like to extend my sincere thanks to Dr. Beibei Wang for her countless suggestions on my research problems, and her contributions in every step of a research project: from literature review, experiment design, and system design to paper drafting. I also thank Prof. Gang Qu, and Prof. Lawrence C. Washington for their support of my Ph.D. defense and their valuable comments.

[1] Calligraphy by Warraich Sahib. CC BY-SA 3.0

My sincere thanks also go to Dr. Chenshu Wu, who has been an amazing mentor throughout this journey, where he always asked new questions and led me to pursue new ideas. His careful approach to new research ideas helped me to justify and rationalize our proposals extensively, and I have learned to keep a bigger picture of the problem in my mind all the time. Without his guidance, it would not have been possible to finish this research.

It has been a great privilege to be a member of the Signal and Information Group (SIG), and I would like to thank my fellow labmates: Dr. Feng Zhang, Dr. Xiaolu Zeng, Dr. Sai Deepika Regani, Dr. Fengyu Wang, Dr. Yuqian Hu, Guozhen Zhu, Wei-Hsiang Wang and Sakila Jayaweera for stimulating discussions, for their support and help during the last four years. This research relies on numerous contributions from each of them and it is impossible to remember all details, as there have been so many.

My special thanks goes to my roommates at our house (WFUP): Dr. Şevket Umut Yürüker, Faizan Wajid, Ömer Akgül and Mustafa Atabey Büyükkaya, for our deep *academic* discussions on Friday nights, which always have been fun and helped me to argue better about my ideas. I would also like to thank friends around the greater DMV area: Ahmet Mehdi Darılmaz, Abdullah Yasir Atalan, Selim Yaman, and Osman Cahit Uğurlu. I would like to extend my thanks to my friends all around the US, although most of them were far away from College Park, their support extended beyond large distances: Burak Varıcı, Mehmet Efe Akengin, Yusuf Dikici, Abdullah Haris Toprak. With these friends and many others, we established an NGO for Turkish students and young professionals in the US, named ONGOR. I will always remember our countless hours of voluntary work for organizing events and supporting other researchers. My research also benefited from the seminars and panels we organized together, and I hope that ONGOR will continue to inspire and help the next generation researchers.

I wish to thank the members of my family-in-law, Yusuf and Zuhal Özer, my sister-in-law Sena and brother-in-law Ömer for being another family for me, and for all the moral support and prayers.

I have always been grateful to my parents Mustafa and Müberra Öztürk for their unconditional love and support in my endeavors. I would like to extend my thanks to my brother Seha, who has been a great brother and researcher, and I continue to learn from him. Furthermore, I would like to thank my youngest brother Ahmet for his support, enthusiasm, and interesting questions about my research, which led me to realize different aspects of my work. Without the support of my family, I would not have made it this far.

Last but certainly not least, my deepest gratitude goes to my wife, Esma, for her love, support and encouragement. Although our paths crossed last year and our lives came together recently, Esma became the sunshine that guided me toward the end of my Ph.D. Her positive attitude, invaluable help and relentless energy made my Ph.D. journey much more enjoyable and easier. It is with heartfelt emotions that I'm finishing this dissertation and looking forward to the new chapter in our life together.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AO	Audio-Only
AoA	Angle of Arrival
AR	Audioradio
AUC	Area Under Curve
CASA	Computational Audiotory Scene Analysis
CER	Character Error Rate
CFAR	Constant False Alarm Rate
CIR	Channel Impulse Response
CNN	Convolutional Neural Network
COTS	Commercial-Off-The-Shelf
CSI	Channel State Information
dB	Decibels
DPRNN	Dual-Path RNN
FMCW	Frequency-Modulated Carrier Wave
FOV	Field of View
GHz	Gigaherts
GPU	Graphics Processing Unit
HHI	Herfindahl-Pirschman Index
HP	High-Pass
Hz	Hertz
IoT	Internet of Things
IQ	In-Phase and Quadrature
irVAD	Interference-Resilient Voice Activity Detection
LDA	Linear-Discriminant Analysis
LLR	Log-Likelihood Ratio
LOS	Line of Sight
LSTM	Long Short-Term Memory

mmWave Millimeter-Wave

NMF Nonnegative Matrix Factorization
NN Neural Network
NLOS Non Line of Sight

PCA Principal Component Analysis
PESQ Perceptual Evaluation of Speech Quality
PIT Permutation-Invariant Training

RF Radio-Frequency
ROC Receiver-Operator Characteristics

SDR Software-Defined Radio
SE Speech Enhancement
SES Speech Enhancement and Separation
SiSDR Scale-Invariant Signal-to-Distortion Ratio
SIR Signal-to-Interference Ratio
SNR Signal-to-Noise Ratio
STFT Short-Time Fourier Transform
STOI Short-Time Intelligibility Metric
SS Speech Separation

UWB Ultra-Wide Band

VAD Voice Activity Detection

WER Word-Error Rate

Chapter 1: Introduction and Motivation

With the proliferation of the Internet of Things (IoT) devices, there are millions of devices that can sense, analyze, and exchange information about the environment; and enable *smart* spaces. These devices are now an integral part of our lives, where the users interact with them using gestures, face, speech, and other biometrics.

Sensing the users or the environment usually requires a variety of sensors, such as microphones, cameras, motion sensors, or PIRs. Among many different sensors, wireless has emerged as a new sensing modality in recent years, thanks to its ubiquity, favorable propagation characteristics, and being a privacy-preserving solution. Within the wireless sensing framework, WiFi has been a very promising medium to monitor environment, thanks to its existence in almost all connected devices. This ubiquity enables to use WiFi connected devices as sensors, with near zero cost. In recent years, WiFi based systems are developed to monitor motion [4], breathing rate [5, 6], sleep [5, 6], speed [7], proximity [8] and localization [9]. These systems usually rely on something called *channel state information (or channel impulse response)*, which can be defined as the echo response of WiFi radio channel. Although the channel state information enables these interesting applications, the performance of WiFi based systems for applications that require precise measurement of the environment has been low. This is due to multiple factors, first, WiFi-based systems lack spatial resolution as WiFi chipsets do not include phased arrays. Signal

processing techniques from phased array processing, such as beamforming are usually not applicable using WiFi. Second, WiFi devices are affected by a variety of noise factors. Although these noise factors do not affect communication significantly, they corrupt the channel state information and limit the usefulness of WiFi for sensing. Because of multiple sources of noise altering the WiFi CSI, it is difficult to extract *sensing* information. Last, WiFi usually has low bandwidth (*e.g.* 20-80 MHz). A wider bandwidth is preferred when the application requires good multipath component separation, (*i.e.*, range resolution). In short, WiFi-based systems have enabled many interesting applications, but additional applications require going beyond the limitations of WiFi.

As WiFi-based sensing has limitations with respect to the spatial and temporal resolution, there have been improvements on multiple fronts. First, millimeter-wave (*a.k.a* extremely high frequency (EHF) for International Communication Union (ITU) designation) devices operating in frequency bands from 30 GHz to 300 GHz have become widely available for industrial and automative applications (*e.g.* [10,11]) and have been used extensively for a variety of applications, such as blind spot monitoring. Second, mmWave bands have become available for both WiFi (e.g. 60GHz WiFi [12]), and 5G networks [13]. Some of the communication chipsets can also be configured to operate as a radar, as in [14]. Third, there is a plethora of research on joint communication and radar devices to enable their coexistence [15]. In short, mmWave radar devices are expected to be part of our smart devices, and they already exist in some [16, 17], as a standalone chip for communication devices.

In recent years, these mmWave devices (*e.g.* [10, 11]) have been used for a multitude of applications. These include vital sign monitoring [18, 19], which includes monitoring of breathing rate and heart rate. Furthermore, in addition to measuring *average* heart rate, more precise information can also be obtained, such as heart rate variability [20] or seismocardiography [21]. Other

applications are contactless gait monitoring and gait based people identification [22–24], remote material identification [14], gesture recognition [25] and vibration sensing [3]. mmWave devices have shown great potential for all these and more applications for environmental monitoring, and have been successfully used in certain commercial products, such as Google Pixel 4 [16].

On the other hand, speech and sound has been one of the most natural ways to interact with the smart devices. To capture the sound, microphones have been employed in smart devices. These microphones are not only used to sense, comprehend, and transmit what the users are saying, but also are capable of inferring information about the environment from the ambient sound. For example, smart speakers like Amazon Alexa can now understand user voices, control IoT devices, and sense particular sounds of interest such as glass breaking or smoke detectors. With these devices becoming widely adopted, new methods to sense acoustic contexts have become even more important. Although microphones are extremely sensitive to capture the *ambient* sound, they have certain limitations.

In this thesis, starting with the observation that mmWave devices have enormous sensing capabilities and microphones have certain limitations, we present a way to sense sound events using an mmWave radio. We use the wireless modality on mmWave bands to extend the sensing capabilities of smart devices, and propose multiple applications to mitigate some fundamental limitations of microphone-based sound sensing. These limitations are usually due to the nature of microphones and some of them are weak source separation, being prone to replay attacks, reduced performance under the presence of severe noise, and not being able to sense beyond walls and acoustically insulated environments. Mitigating these problems would enable further capabilities, such as sustained performance over noisy environments, increased privacy and security capabilities, and sound-awareness of outside environments. Thanks to the different channel

and propagation characteristics, and processing capabilities of mmWave devices, we show that these limitations can be overcome, and a new era of smart devices with additional sound sensing capabilities is possible.

With these applications in mind, we first establish the theory to sense sound-induced vibration from ambient objects, such as a piece of aluminum foil, which is possible thanks to higher bandwidth, and shorter wavelength of mmWave radio devices. Next, we propose a system to detect, localize and separate multiple sound sources. The proposed system includes modules to denoise, and compensate for differences in the received signal arising from the different medium of sensing. Afterward, we investigate the feasibility of multiple problems using a radio-based or a multimodal (*i.e.* audioradio)[1] approach, such as speaker verification, speech enhancement and separation, and speech activity detection.

More specifically, in Chapter 2, we propose a method to extract sound-induced vibration. Since the vibration on object surfaces due to sound are on the orders of micrometers [1], high fidelity approach is needed to extract speech. To that end, we show how our method outperforms existing methods in the literature, and we present a material study to understand practical limits on sound reconstruction from the *environment*, using a commercially available of-the-shelf (COTS) device.

In Chapter 3, we propose a method to detect, and localize sound-induced vibration signals, in the presence of background noise and vibrations. Inspired by the recent advances in deep learning and speech denoising, we propose a data-driven system to combat the effects of the channel and show the effectiveness of our proposal.

[1]Inspired by the word *audiovisual* we conjoin the terms *audio* and *radio* to refer a multimodal system that comprises of both modalities.

In Chapter 4, we propose a multimodal method to first detect and localize speech sources, and then use the radio side-channel information in tandem with a microphone to achieve speech enhancement and separation at high fidelity. Introduction of radio modality helps to remove some of the inherent problems in acoustics domain, and we illustrate the benefits of our proposal extensively.

In Chapter 5, we use the radio-channel information to detect voice activity of users in the field of view (FoV) of the radar. Inherent to the radio-channel, our proposal is robust against background noise and interference, and can detect voice activity of multiple users individually. Our experiments demonstrate that the vibration of the vocal folds is distinctive enough to be separated from the motion of the targets as well, and a radio-based system is robust against different types of motion.

In Chapter 6, we provide a brief summary of the thesis and give future research directions for each system in the preceding chapters. Furthermore, we provide an insight into the future of *wireless* sensing, and conclude with our final remarks.

1.1 Related Work

This dissertation is at the intersection of acoustic and wireless sensing, specifically with applications to ambient vibration monitoring, speech enhancement and separation, and voice activity detection. The related literature to wireless and speech sensing is given in 1.1.1 where we introduce several applications of wireless-based sensing and focus on sound-related tasks. In addition, we briefly present other modalities to sense sound events, such as ultrasound, or a high-speed camera. In Section 1.1.2, we present the speech enhancement and separation methods

together. In Section 1.1.3, we provide the related framework to voice (speech) activity detection.

1.1.1 Wireless and Speech Sensing

As a new medium for sensing, wireless, has been an emerging field [26–34] in recent years. Wireless-based systems have a variety of applications such as vital signs monitoring (*e.g.* breathing, heart-rate) [35–37], gesture recognition [38], indoor localization [39], motion detection [4], or gait recognition [22,24,40]. Wireless signals are reflected from the objects in the environment, and these reflections contain information about the environment.

Since high-frequency signals are more sensitive to changes in the environment, due to shorter wavelength, mmWave wireless devices are shown to capture minute changes in the environment. This minute motion can be due to the motion of lips, vocal folds, or sound pressure on object surfaces. The theory to explain the relationship between sound induced vibration and radio signals has been established in [41]. In capturing *sound* using a radio device, two different approaches have been used, passive or active. Active approaches focus on capturing the *source* vibration, such as from human vocal folds, guitar strings, or speaker diaphragms. On the other hand, passive sources are objects in the environment that vibrate passively, due to another active source. Due to the amount of displacement (*e.g.* 100-200μm in vocal folds vibration compared to 20μm on a piece of aluminum foil), the performance for the active sources has been higher. There has been relatively minimal work on sensing sound from passive sources. Some introductory works [42–44], show that it is possible to recover bandlimited sound from objects in the environment. Moreover, existing works mainly focus on lower frequency signals and do not address the fundamental limitations of high frequency. Our work [45] explores how to overcome those

limitations. On the other hand, there has been more research activity to sense active sources using radio signals, and their applications. Vocal folds vibration can be captured remotely [46, 47]. Using deep learning, speech can be reconstructed from human throats [48] using a customized 16x16 hardware operating at 24 GHz frequency bands. Furthermore, mmWave *signature* of the vocal folds can be used for voice liveness detection and verification in [45, 49], or for speaker identification [50]. Furthermore, acoustics and mmWave signals are combined further for speech recognition [51], and for speech enhancement and separation in our work [52, 53]. Last, radars have also been used for capturing vibration characteristics of loudspeakers in [54], and that of machinery in [3].

It is also possible to capture sound and speech using low-frequency signals with some additional constraints. As an example, a pioneer work [55] uses a 2.4 GHz SDR to capture sound. Likewise, WiHear [56] captures signatures of lip motions with WiFi, and matches different sounds with those in a limited dictionary. Other radio-frequency based methods include contact-based RF sensors [57, 58] for silent speech reconstruction, and ultrawideband devices for sound reconstruction from speaker diaphragms [2].

1.1.1.1 Ultrasound-Based Speech Sensing

Ultrasound is another modality that allows to sense environmental changes. Various ultrasound-based methods to sense live speech from human bodies have been proposed by the literature. As ultrasound signals cannot penetrate the human skin, the main focus of these approaches has been capturing lip motion. These works capture lip motion [59], recognize the speaker [60], synthesize speech [61], or enhance sound quality [62, 63]. These approaches usually have a very limited

range, and require a prior dictionary, as the motion cannot be related to sound immediately.

1.1.1.2 Light-Based Speech Sensing

VisualMic [1] recovers sound from the *environment* (*e.g.* bag of chips) using high-speed cameras. A similar phenomenon is exploited using a telescope to measure changes in light intensity of a lamp in [64]. Laser has been famously used for eavesdropping on different environments through vibration induced displacements on windows [65]. Furthermore, lidar [66], and depth cameras [67] can also be used for sensing sound. Although a variety of modalities or sensors in *vision* domain enables sound sensing, these methods usually require expensive specialized hardware. Developing a radio-based sound sensing system would be more low-cost, can penetrate through objects, and can work in the dark.

1.1.1.3 IMU-Based Speech Sensing

Other sensors have been used for sensing sound or speech, such as accelerometer [68], gyroscope [69], or vibration motor [70]. AccelWord [68] uses the accelerometer of a mobile device to sense a limited dictionary, whereas VibroPhone [70] uses a reversely connected vibration motor of a mobile device to sense acoustic changes at the *destination*. GyroPhone [69] illustrates the feasibility of basic speaker recognition using a gyroscope. All these methods sense the sound at *destination* like contact microphones, and has similar drawbacks to microphones, in addition to their limited bandwidth.

1.1.2 Speech Enhancement and Separation

In this subsection, we review a multitude of methods for speech enhancement and separation. We first review the audio-based speech enhancement and separation methods in 1.1.2.1 and investigate multimodal systems in 1.1.2.2.

1.1.2.1 Microphone-Based Methods

Early methods in speech enhancement and separation include mathematical modeling based approaches. One of the traditional methods is computational auditory scene analysis (CASA) [71] which models human perception to solve scene analysis problem. This approach can be combined with a variety of methods, such as pitch-estimation based separation in [72]. Another method is nonnegative matrix factorization (NMF) [73], where usually the speech (and potentially noise and music) signals are represented with respect to a set of basis vectors, and the mixture spectrograms are decomposed into its components. Other approaches include probabilistic methods such as [74]. In this approach, some feature vectors (such as mel frequency cepstral coefficients (MFCC)) are extracted and modeled as a random variable with a distribution. By modeling the time series data of speech and noise signals as a Hidden Markov Model, both sequences are decoded using a maximum likelihood estimation method. On the other hand, these methods usually cannot generalize well to unseen speakers [75], which is a major limiting factor for their performance.

Deep learning based methods outperformed classical approaches recently [75]. Early deep learning based methods take the input mixture as input, and estimated the clean signal spectrograms. More recently, instead of estimating output representation directly, deep learning-based

methods usually estimate a mask that is multiplied by the input. Since the input spectrograms are complex, and masking with complex numbers is not straightforward, different mask estimation methods have been investigated. Early works estimated a binary magnitude mask [76] to modify the noisy magnitude spectrograms. In this method, estimated clean spectrogram amplitudes are used with the noisy input spectrogram phase, as the importance of phase is relatively lower compared to the magnitude [77]. Since the performance of magnitude-only mask estimation is limited, researchers proposed other masks, such as STFT spectral mask [78], and complex ratio mask [79]. On the other hand, some approaches, such as PHASEN [80], estimate amplitude and phase masks separately in the spectrogram domain to enable real-valued operations for neural networks. Extensions of real-valued neural networks, such as complex neural networks in DCCRN [81], can also be used for speech enhancement. In this approach, the authors propose complex valued CNNs and define complex CNN operations to enable spectrogram estimation.

All these given approaches use spectrogram representation of the speech signals, which is usually a good time-frequency representation. As noted by researchers [75], modifying spectrograms usually create musical artifacts and are not very plausible. To solve this, different approaches included time-domain processing. As an example, in SEGAN [82], a time-domain speech enhancement system using generative adversarial networks has been proposed. On the other hand, *learnable* time-frequency representations are also explored. ConvTasNet [83] show a breakthrough by performing better than ideal ratio mask for SS, with an adaptive/learnable encoder, instead of classical STFT. After ConvTasNet, many different works adopted learnable encoders, and modified ConvTasNet structure for increased performance. As an example, the fully convolutional layers in [83] are replaced by dual-path RNN (DPRNN-TasNet) [84], dual-path transformer network (DPTNET) [85], and fully attention layers in SepFormer [86].

Source association and tracking problems can be solved with frame-level PIT [87] and utterance-level PIT [88]. Even though these methods mitigate the problem, and estimate the same speaker's speech for a given frame, they can fail when the speakers have similar pitch and speaking characteristics [89]. The number of sources can be estimated by deep clustering [90] or deep attractor networks [91]. However, these models still have the source tracking problem over long time, which is started to be addressed recently [92].

1.1.2.2 Multimodal Methods

Vision-based works use different features as input, such as face embeddings [93], lip embeddings [94] or optical flow [89]. These methods use STFT representation, although time-domain processing is also possible [95]. [96] estimates faces from the speech signals, whereas [97] uses picture of a speaker for separation. Audiovisual methods have complex processing pipelines, require good lighting, and raise privacy concerns.

On the other hand, ultrasound can also be used for speech generation [61], speaker recognition [60], and speech enhancement [62]. UltraSE [63] uses a deep learning to enhance single speaker signals. Ultrasound signals can only work at a short range (*e.g.* 15cm), and are too coarse to measure fine-grained vocal folds vibration [63].

1.1.3 Voice Activity Detection

Although voice activity detection is mostly an acoustics domain problem, different modalities are also explored by researchers in previous years. We first provide acoustics based VAD systems in Section 1.1.3.1, and present vision and ultrasound based approaches in Section 1.1.3.2

and 1.1.3.3.

1.1.3.1 Acoustic-based VAD

Early VAD algorithms relied on the statistical difference between speech and background noise [98]. This approach extracts discrete Fourier transform (DFT) coefficients from the time signals, and models speech and noise as independent Gaussian random variables. After estimating the probabilities of the speech and noise, the algorithm classifies the signal based on the likelihood ratio of the two probabilities. Several improvements to this method has been proposed, such as replacing the likelihood ratio with smoothed likelihood ratio in [99], or modeling the speech as a Laplacian random variable in [100]. On the other hand, the performance of these approaches have been limited, due to the diversity of noise types in real-world scenarios.

In recent years, statistical models are replaced by deep learning-based methods, as in such as deep belief networks in [101], recurrent neural networks in [102], and outperformed existing statistical modeling based methods. Although these methods work fairly well in *in domain* settings, their performance decrease in unseen locations. Authors in [103] proposed a method to improve robustness. On the other hand, unless there is some prior information, such as speaker embeddings or spatial separation with microphone arrays, these methods fail to focus on a particular user and not useful for *irVAD*. To solve *speaker targeted VAD* problem, authors in [104] proposed to extract the speaker embeddings of the users *a priori*, and use this information to selectively detect the voice activity of a particular user. They have extended their work in [105], by accompanying training-free setting in the system. The evaluation is done in noisy environments, either with background speech or target speech (*i.e.* there is no overlap between target and

interfering speakers), and therefore lacks some practical considerations.

A closely related topic to voice activity detection in multiple speaker environments is speaker diarization (*a.k.a who spoke when*) problem [106]. In speaker diarization, meeting recordings with multiple participants with potentially overlapping speech are considered. Although it is relevant, speaker diarization requires identification of speakers, and assignment of labeling, and usually involve computationally complex deep networks [107].

1.1.3.2 Vision Based VAD

Visual stream has also been used to detect voice activity, as the lip motion is an indicator of speech [108, 109]. Early works use classical feature extraction algorithms, such as PCA/LDA [110], whereas more recent works explore deep learning-based feature extraction from videos [111] or from dynamic images [112]. In [113], a multimodal VAD that can work robustly in challenging environments is proposed, by using WaveNet encoder and residual networks. Even for audiovisual systems, annotation is a difficult task, where annotation tools are developed in [114, 115], or automatic annotation is explored in [111].

These works require perfect visibility of lip and face area, good lighting conditions, and potentially raise privacy concerns. Furthermore, the granularity of speech detection is based on the frame-level (*e.g.*at 24 fps, detection can be made for frames of 42ms), and usually require *several data frames*, which is not very practical in real-time applications.

1.1.3.3 Ultrasound Based VAD

As a medium that has *spatial* sensing capability, ultrasound has also been explored for VAD. Early works, such as [116, 117], propose an ultrasound-based frame-level VAD and evaluate the performance in controlled settings. In [118], the authors used off-the-shelf microphones and speakers to achieve a similar goal at a higher accuracy. In recent years, ultrasound channel has also been combined with deep learning, such as [119].

1.2 Dissertation Outline and Contributions

Based on the limitations of the existing sound sensing, enhancement, and voice activity detection systems, we have demonstrated the need and importance of enriching those applications with a remote, contactless, robust, and potentially ubiquitous sensing modality, wireless.

In this dissertation, we eliminate some of these limitations using mmWave-based sensing, and propose multiple methods to address different applications. Our unimodal and multimodal systems address the problem of *spatial* sound sensing, extracting sound from their *source* and using radio signals as a side channel for speech activity detection, enhancement, and sensing. We briefly introduce these applications in the following sections, and elaborate on those further in the corresponding sections.

1.2.1 Sound Sensing Preliminaries (Chapter II)

In Chapter 2, we model the vibration on object surfaces due to sound for mmWave devices. We propose a method for the recovery of sound and conduct experiments with various materials to investigate the feasibility of sound reconstruction. We further evaluate the effect of distance

and placement to understand the practical limits on the sound reconstruction. The results show that, by using a commodity off-the-shelf radar, it is possible to capture a significant amount of sound from the environment.

1.2.2 *RadioMic*: mmWave-Based Sound Sensing System (Chapter III)

In Chapter 3, we propose a radio-based sound sensing system, named *RadioMic*. Mainly, we focus on the fundamental issues of microphones, such as weak source separation, limited range in the presence of acoustic insulation, and being prone to multiple side-channel attacks.

Since voice interfaces relying on microphones to sense sound have become an integral part of our lives, these problems have become even more important. *RadioMic* is a radio-based sound sensing system to mitigate these issues and enrich sound applications. *RadioMic* constructs sound based on tiny vibrations on active sources (*e.g.*, a speaker diaphragm) or object surfaces (*e.g.*, paper bag), and can work through walls, even a soundproof one. To convert the extremely weak sound vibration in the radio signals into sound signals, *RadioMic* introduces *radio acoustics*, and presents training-free approaches for robust sound detection and high-fidelity sound recovery. It then exploits a neural network to further enhance the recovered sound by expanding the recoverable frequencies and reducing the noises. *RadioMic* translates massive online audios to synthesized data to train the network and thus minimizes the need for radio-frequency (RF) data. We thoroughly evaluate different components of *RadioMic* under different scenarios using a commodity mmWave radar. The results show *RadioMic* outperforms the state-of-the-art systems significantly. We believe *RadioMic* provides new horizons for sound sensing and inspires attractive sensing capabilities of mmWave sensing devices.

1.2.3 *RadioSES*: Sound Enhancement and Separation System (Chapter IV)

In Chapter 4, we aim to solve another major limitation of the microphone-based systems, namely, being prone to interference and noise. Speech enhancement and separation have been a long-standing problem, as there have been recent advances using a single microphone. Although microphones perform well in constrained settings, their performance for speech separation decreases in noisy conditions. In Chapter 4, we propose *RadioSES*, an audioradio speech enhancement and separation system that overcomes inherent problems in audio-only systems. By fusing a complementary radio modality, *RadioSES* can estimate the number of speakers, solve the source association problem, separate and enhance noisy mixture speeches, and improve both intelligibility and perceptual quality. We perform millimeter-wave sensing to detect and localize speakers, and introduce an audioradio deep learning framework to fuse the separate radio features with the mixed audio features. Extensive experiments using commercial off-the-shelf devices show that *RadioSES* outperforms a variety of state-of-the-art baselines, with consistent performance gains in different environmental settings. Compared with the audiovisual methods, *RadioSES* provides similar improvements (*e.g.* 3 dB gains in SiSDR), along with the benefits of lower computational complexity and less privacy concerning.

1.2.4 *RadioVAD*: Voice Activity Detection System (Chapter V)

In Chapter 5, we introduce a mmWave-based voice activity detection system to address noise and interference robustness problem of microphones. Because of this, microphone-based voice activity detection systems usually require hotword detection and they cannot perform well under the presence of interference and noise. Users attending online meetings in noisy envi-

ronments usually mute and unmute their microphones manually due to the limited performance of interference-resilient VAD. In order to *automate* voice detection in challenging environments without dictionary limitations, we explore beyond microphones and propose *RadioVAD*, which is an mmWave based voice activity detection system. Our extensive experiments in multiple places with several users indicate that mmWave-based VAD can match and surpass the performance of an audio-based VAD in noisy conditions, while being robust against interference.

Chapter 2: Primer on Sensing Sound Induced Vibration

2.1 Introduction

Microphones have been used to sense acoustic events in the environment for many application since their introduction. Similar to the human auditory system that detects tiny changes in the air pressure, where the sensing is done at the *destination* (*i.e.* ears), IoT-based voice interfaces also capture the air pressure at the *destination* via microphones.

In recent years, other modalities to sense the sound have been proposed, such as laser microphone [65], visual (camera) microphone [1], vibration motor (VibroPhone) [70], accelerometer-based hotword detection [68], gyroscope [69] and light-based [64] methods. Laser microphones [65] generally use a secondary medium such as windows of a building to capture changes in the phase of light due to sound induced vibration. Similarly, visual microphone [1] and lamphone [64] enables to *sense* the air pressure at intermediary objects in a passive way, which we name as sound sensing from the *environment*. On the other hand, accelerometer-based methods capture sound from the *source*, whereas VibroPhone [70] requires to capture air vibration at the sensor, or at the *destination*.

Sound sensing from the *environment* brings advantages of remote sensing and eavesdropping, yet previous methods require line-of-sight operation [1, 65] or ambient light [1], due to the modality of choice. This chapter explains sensing sound from the *environment* by using

an mmWave radar, in order to enable non-line-of-sight operation, and potentially capture sound through sound-proof materials, as the propagation characteristics of acoustic and radio signals differ. Other advantages of using a mmWave radar to sense sound will be presented in Chapter 3.

There are many challenges to sense sound with an mmWave device. First, radars sense displacement on objects by measuring changes in the phase of the returned signal. Although mmWave radars have much shorter wavelengths compared to conventional wireless devices, vibration displacement due to sound are orders of magnitude smaller, on the order of several micrometers. Therefore, it is extremely challenging to capture the changes in the phase, especially in the presence of noise. In addition, radars cannot differentiate reflections from nearby objects and cannot *focus* on a particular object, without creating narrow beams either with high-resolution beamforming or highly directional antennas.

In what follows, we model the vibration of an object and establish the mathematical relationship between the object vibration and radar signal. Although our vibration model is similar to that of [3, 18, 120], our approach is different, as the sound-induced vibration is orders of magnitude weaker than the mechanical vibrations considered in those works. Our model uses sound vibration on objects to construct the audio signal, which is affected by the frequency response of the acoustic vibration, and the radio channel between the object and the radar, which will be explained later. Next, we evaluate vibration response of various materials and effects of distance, orientation, and speaker distance, and show that it is feasible to capture sound-induced vibrations by using a COTS radar.

2.2 Radar Preliminaries

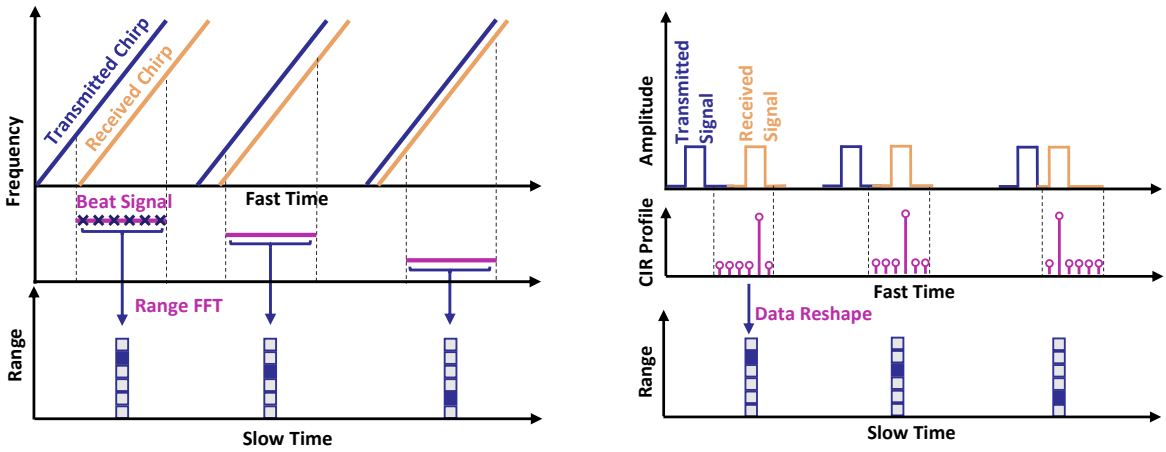
In this section, we present the preliminaries with respect to sound sensing using radar. We first define the channel impulse response for an *impulse response* radar device, and introduce an equivalent representation for an *frequency-modulated carrier wave (FMCW)* radar.

An FMCW radar transmits a single tone signal with linearly increasing frequency, called a chirp, and captures the echoes from the environment. Time delay of the echoes could be extracted by calculating the amount of frequency shift between the transmitted and received signals, which can be converted to a range information. This range information is used to differentiate an object from the other reflectors in the environment. In order to obtain the range information, the frequency shift between transmitted and received signals are calculated by applying FFT, which is usually known as *Range-FFT* [121]. The output of Range-FFT can be considered as CIR, $g(t, \tau)$. On the other hand, an impulse radar transmits short impulses, such as Savitzky-Golay sequences, and uses correlation (*e.g.*match filter) at the receiver to extract CIR. Two processing mechanisms are illustrated in Fig. 2.1

In general, CIR of an RF signal can be given as

$$g(t, \tau) = \sum_{l=0}^{L-1} \alpha_l(t)\delta(\tau - \tau_l(t)), \tag{2.1}$$

where t and τ refers to long and short time, L denotes the number of range bins (sampling wrt. distance), α_l denotes complex scaling factor, τ_l is the roundtrip duration from range bin l, and $\delta(\cdot)$ represents Dirac delta function, indicating the presence of an object. Assuming no multipath, and an object of interest at range bin l^*, corresponding to the time delay τ^*, CIR of that range bin



(a) FMCW Radar Processing

(b) Impulse Radar Processing

Figure 2.1: Radar Processing Mechanisms

can be given as:

$$g(t, \tau^*) = \alpha_{l^*}(t) \exp(-j2\pi f_c \tau_l^*(t)), \tag{2.2}$$

where f_c denotes the carrier frequency. If we assume the object to remain stationary in range bin l^*, we can drop the variables τ^*, and l^*, convert time delay into range, and rewrite the CIR as:

$$g(t) = \alpha(t) \exp(-j2\pi R(t)/\lambda)), \tag{2.3}$$

where $R(t)$ denotes the actual distance of the object, and λ denotes the wavelength.

Using CIR (or converted FMCW waveform), *range-Doppler spectrogram* can be extracted by a short-time Fourier Transform (STFT) operation. STFT is FFT operations applied in the t dimension in $g(t, \tau)$ for subsets of long-time indices, called *frames*, to ensure capturing temporal changes, where the samples are multiplied with a window function prior to the transform. We denote the output range-Doppler spectrograms as $G(f, r, k)$, where $f \in (-N_s/2, N_s/2)$ denotes frequency shift, r corresponds to range bins (equivalent to τ_l), and k is the frame index. We note

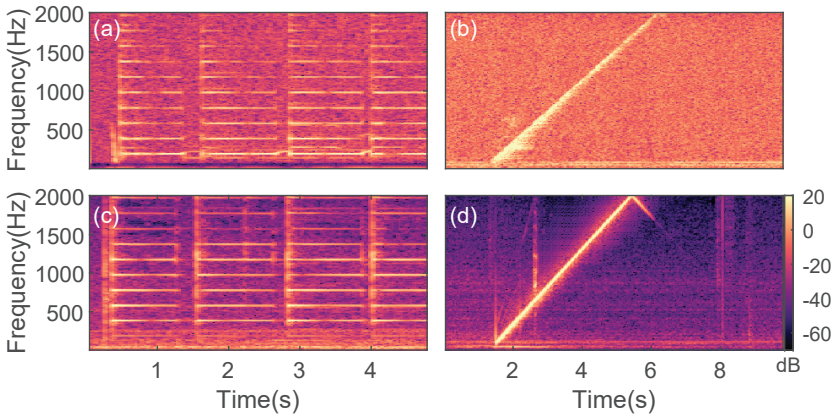


Figure 2.2: Examples of radio-sensed sound. Top row (a,b) are radio reconstruction, bottom (c,d) are microphone references. Left: guitar sound (Note G3, 196 Hz); Right: frequency sweep sound sensed from aluminum foil.

that G is defined for both positive and negative frequencies, corresponding to different motion directions of the objects, which will be used in this manuscript extensively. Denoting short-time indices with $n_k(i) \triangleq (k-1)(N_f-N_{ov})+i$ for $i \in \{1, ..., N_f-1\}$, where N_{ov}, N_f amount of overlap and frame length in terms of number of samples. Consequently, range-Doppler spectrogram is defined as

$$G(f, r, k) = \left| \sum_{i=0}^{N_f-1} W(i) g(r, n_k(i)) \exp\left(j \frac{2\pi i T f}{N_f}\right) \right|^2, \tag{2.4}$$

where $W(i)$ represents a finite length windowing function used to fine tune the resolution between time and frequency domains.

2.3 Sound Mechanics

Sound is basically modulation of medium[1] pressure through various mechanisms. It is generated by a vibrating surface, and the modulation signal travels through in-place motion of air

[1] Without loss of generality, we assume the medium to be air throughout this dissertation.

molecules. A vibrating surface could be a speaker diaphragm, human throat, strings of musical instrument such as a guitar, and many daily objects like a paper bag. In the case of speakers, motion on the speaker diaphragm modulates the signal, whereas in human throat, vocal cords create the vibration, with the mouth and lips operating as additional filters, based on the source-filter model [122]. To sense the sound, the same mechanism is employed at the microphones to convert the changes in the air pressure into electrical signal, via suitable diaphragms and electrical circuitry. Microphone diaphragms are designed to be sensitive to air vibration and optimized to capture the range of audible frequencies (20Hz-2kHz), and even beyond [123].

Vibration (motion) on object surfaces is proportional to the transmitted energy of sound from the air to the object and depends on multiple factors, such as inertia and signal frequency [124]. Denoting the acoustic signal with $a(t)$ we can model the sound-induced displacement as:

$$x(t) = h \star a(t), \tag{2.5}$$

where h denotes the vibration generation mechanism for an active source or the impulse response of the air-to-object interface for a passive object, and $\star$ represents convolution.

Using the receiver model in Chapter 2 for a vibrating object (*i.e.*, sound source) at distance $R(t)$, we can decompose the range value into the static and vibrating part as $R(t) = R_0 + x(t)$. As can be seen, there is a direct relationship between the CIR $g(t)$ and the phase of the returned signal. By extracting the phase, $g(t)$ could be used to derive $R(t)$, and therefore, the vibration signal, $x(t)$. We further omit the temporal dependency of α, as we assume the object to be stationary, and the effect of displacement due to vibration on path loss to be negligible.

So far, we have assumed to have the vibrating object in the line of the radar solely, and

did not account for other reflections from the environment. As suggested by (2.2), $g(t)$ lies on a circle in the IQ plane with center at the origin. However, due to various background reflections, $g(t)$ is actually superimposed with a background vector, and the circle center is shifted from the origin. Thus, $g(t)$ can be written as:

$$g(t) = \alpha \exp\left(-j2\pi \frac{R(t)}{\lambda}\right) + \alpha_B(t) \exp(j\gamma(t)) + w(t), \tag{2.6}$$

where $\alpha_B(t)$ and $\gamma(t)$ are the amplitude and phase shift caused by the sum of all background reflections and vibrations, and $w(t)$ is the additive white noise term.

2.4 Signal Extraction

Using the signal model in (2.6), the acoustic signal is recovered by first filtering out the interference and background and approximating the remaining signal with a line fit to further reduce noise.

We first apply an FIR high-pass filter as the changes in the background usually have much lower frequencies. The resulting signal, $\hat{y}(t)$ can be given as:

$$\hat{g}(t) \approx \alpha \exp\left(-j2\pi \frac{R(t)}{\lambda}\right) - \alpha \exp(j\gamma_R) + \hat{w}(t), \tag{2.7}$$

where $\hat{w}(t)$ is the filtered noise term, and $\alpha \exp(j\gamma_R) \approx \alpha \exp(-j2\pi \frac{R_0}{\lambda})$ is the center of a circle in IQ plane, which is due to the background vector. The signal component, $\exp(-jg\pi \frac{R(t)}{\lambda})$, remains mostly unchanged, due to the frequencies of interest with sound signals and this operation moves the arc of the vibration signal to the origin in the IQ plane. Furthermore, this operation reduces

any long-term drifting in IQ plane, caused by the hardware.

As explained in the previous section the curvature of the arc is in the order of 1° for μm displacement with a mmWave device, by projecting the arc, $\alpha \exp(-j2\pi \frac{R(t)}{\lambda})$, onto the tangent line at $\alpha \exp(-j2\pi \frac{R_0}{\lambda})$, we can remove the background and approximate $\hat{g}(t)$ as

$$\hat{g}(t) \approx nx(t) + \hat{w}(t), \tag{2.8}$$

where $n = \alpha \frac{-2\pi}{\lambda} \exp(-j(\frac{\pi}{2} + 2\pi \frac{R_0}{\lambda}))$.

Eq. (2.8) suggests that $\hat{g}(t)$ already has the real-valued sound signal $x(t)$, scaled and projected in complex plane, with an additional noise. Using the fact that the projection onto an arbitrary line does not change noise variance, we can estimate $x(t)$ with minimum mean squared error (MMSE) criteria. The estimate is given as:

$$\hat{x}(t) = \mathcal{R}\{\hat{g}(t) \exp(-j\hat{\theta})\}, \tag{2.9}$$

where $\mathcal{R}$ is the real value operator. Angle $\hat{\theta}$ can be found as:

$$\hat{\theta} = \arg\min_{\theta} \|\hat{g}(t) - \mathcal{R}\{\hat{g}(t) \exp(-j\theta)\} \exp(j\theta)\|^2. \tag{2.10}$$

2.5 Experiments

As our system tries to capture the sound signal from an intermediary object, its performance depends on two physical phenomena, (1) the vibration properties of the object, which is affected by the speaker location and object material, and (2) the radar response, which can change due

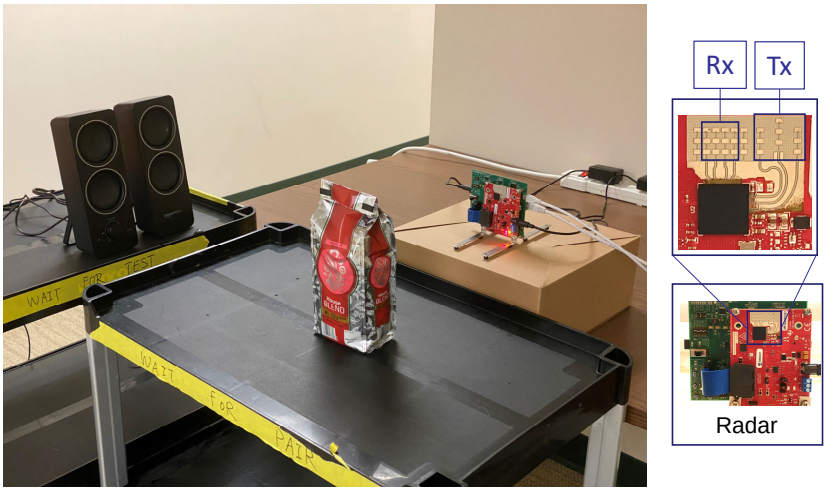
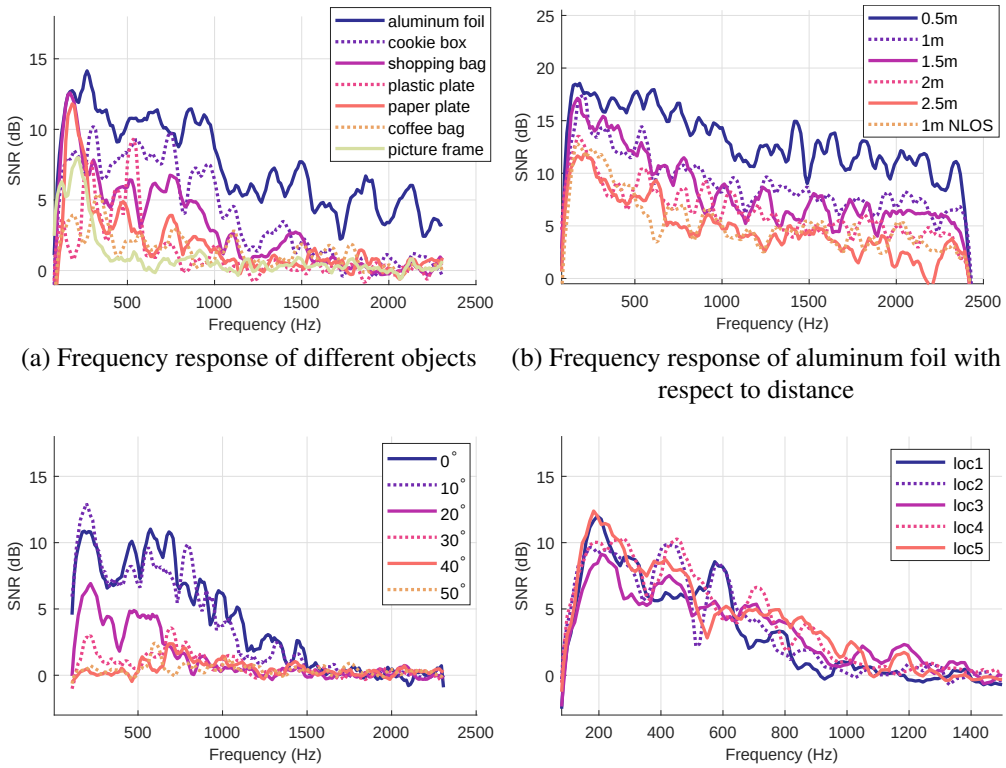


Figure 2.3: Experimental Setting

to various factors, such as distance, reflectivity, orientation, and specularity. In this subsection, instead of evaluating performance with arbitrary sound files, we evaluate the impacts of the above factors on the frequency responses, as this can be used to estimate the waveforms of output sound files. Moreover, these responses can be used to investigate the frequency dependency of vibration and radar channel between the object and the device.

As shown in Fig. 2.3, we place the object and source at different distances. We play a frequency sweep through speakers from 100 Hz to 3200 Hz, and capture the vibration amplitude at corresponding frequencies. The sound pressure at 300 Hz is measured to be 86 dB, (0.4Pa), at 0.5m away from the speakers. In addition, by estimating the noise variance during time instances of silence, we present the signal-to-noise ratio (SNR) in each experiment. We present two evaluation cases, object material, and distance with respect to the radio device, speaker location and object orientation.



(a) Frequency response of different objects

(b) Frequency response of aluminum foil with respect to distance

(c) Frequency response of aluminum foil with respect to angle

(d) Frequency response at different speaker locations

Figure 2.4: Frequency Response Experiments

2.5.1 Object Material

In this experiment, we investigate the frequency responses of various materials. Under the setting shown in Fig. 2.3, we evaluate frequency responses of cookie box, facial mask box (paper, 8×15 cm), aluminum foil (30×30 cm), paper and plastic plates ($r = 6cm$), cartoon bag and picture frame in Fig. 2.4a. We also test several circular shaped objects (*e.g.* metallic thermos, ceramic mug), yet we do not report their results, as almost no signal can be recovered. These results indicate that given that the object is placed in front of the radar, it is possible to recover significant sound content with aluminum foil and varieties of papers. As the thickness and inertia of the objects increase, the amplitudes in the frequency response decrease.

2.5.2 Object Distance

In this experiment, we evaluate the effect of distance by placing an aluminum foil at varying distances. In order to keep the sound levels at the object surface the same, we keep the object to speaker distance as 0.5m. We place the speakers behind the object to ensure separation, which improves the overall performance slightly compared to the previous experiment. As shown in Fig. 2.4b, the SNR reduces with an increasing distance; but it is still possible to capture some sound content up to 2.5 meters. It should be noted that, when the object is very close, the reconstructed signal quality increases significantly, possibly due to the surface area within the beamwidth. Lastly, we also test an NLOS setting by putting a wooden frame in between, and placed the foil at 1m. As seen, through-wall setting reduces the SNR significantly, but it should still be possible to reconstruct frequencies up to 1 kHz, with enough amplitude.

2.5.3 Speaker Location

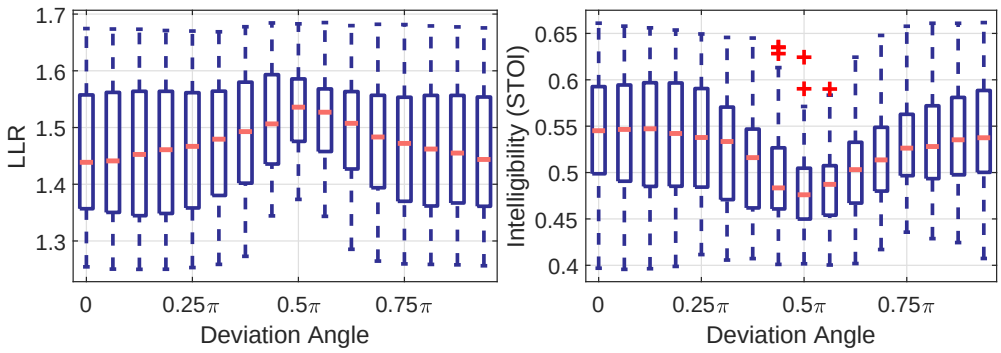
In this experiment, we investigate the effect of varying speaker location. In a 4m $\times$ 7m office room with objects such as desks, chairs, and computer equipment, we place the speaker in five different locations and capture the frequency response. We place the radar and object in the center of the room, and vary the location of the speaker, closest location being 2.5m away from the object, and furthest being 3.5m away. Furthermore, we change the orientation of speakers in this experiment, and they do not face the object, with an acute angle of at least 90°. As seen in Fig. 2.4d, the relationship between the speaker location and the frequency response is weak, and our method indeed captures ambient sound, instead of some vibration leakage through the ground or electromagnetic radiation from the speakers.

2.5.4 Object Orientation

In this experiment, we change the orientation of the object, which affects the response by changing the direction of vibration, as well as the strength of the returned signal, due to changes in the surface area facing the radar. As the previous experiment suggests, ambient reflections cause significant vibration content; and in order to reduce the reflections from walls, we use an anechoic chamber for this experiment, and place the radar at 0.75m. Moreover, we keep the distance and angle between the object and the speaker the same, and change the angle between the radar and the object with 10° increments. As shown in Fig. 2.4c, it is possible to capture a significant amount of vibration with an angle up to 20°, whereas increasing the angle any further gradually reduces the returned signal amplitude. This indicates that, unlike lasers, an mmWave-based system can tolerate changes in the angle up to a certain extent. The overall response in this experiment is weaker, due to the sound absorbing materials in the room, and at 300 Hz the sound pressure on the object surface is measured to be 81 dB.

2.5.5 Verification of Line-Projection

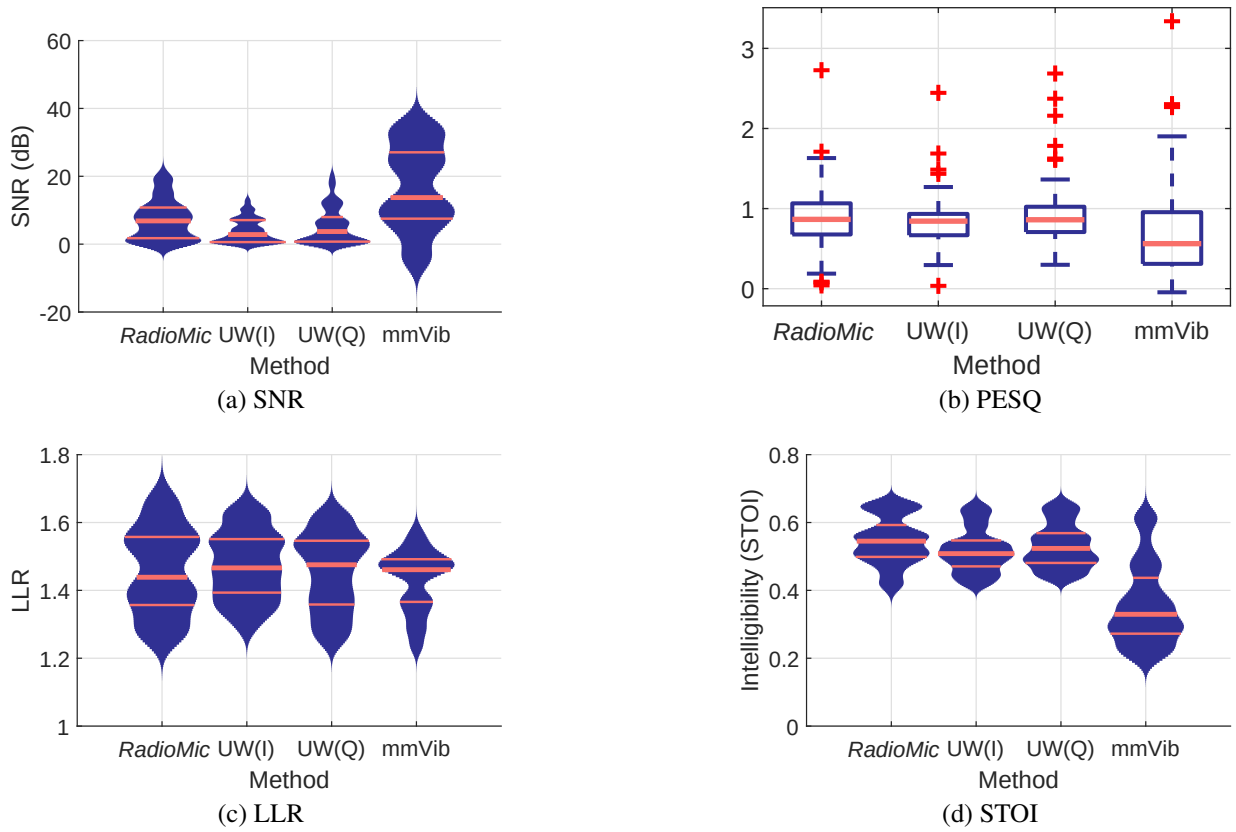
Before concluding this chapter, we investigate the optimality of the signal extraction approach proposed in Section 2.4 by two different evaluation cases. First, we use different projection angles to show how a significant drop is observed when there is 90° deviation from the optimal angle, which can be seen in Fig. 2.5. As can be seen, the optimal performance can be achieved when the angle of projection is used (0°). Furthermore, 90° results in the worst performance with respect to both metrics, which is the orthogonal axis to the optimal axis and is expected. Second, we compare the results with existing methods in the literature. Comparison



(a) LLR (b) STOI

Figure 2.5: Verification of line projection

of our extraction method with other methods, such as [2] and [3] is given in Fig. 2.6. As can be seen, our method outperforms other methods in terms of PESQ [125], STOI [126] and LLR [127]. These metrics will be further explained and investigated in Chapter 3.



(a) SNR (b) PESQ

(c) LLR (d) STOI

Figure 2.6: Comparison with UWHear [2] and mmVib [3]. UW(I) and UW(Q) denotes in-phase and quadrature signals extracted by UWHear [2] respectively. Horizontal lines on violin plots represent 25th, 50th and 75th percentile, respectively. Box plot is used for (b) due to outliers.

Chapter 3: *RadioMic*: mmWave-based Sound Sensing System

3.1 Introduction

In Chapter 2, we introduced preliminaries to extract sound-induced vibration from object surfaces. Now, we shift our focus to a comprehensive system that can detect, and localize sound sources, and mitigate for the changes caused by the channel. We explain how a system, named *RadioMic*, can achieve all of these tasks and enable various applications, some of which are illustrated in Fig. 3.1. *RadioMic* can detect, recover and classify sound from sources in multiple environments. It can recover various types of sounds, such as music, speech, and environmental sound, from both *active sources* (*e.g.*speakers or human throats) and *passive sources* (*e.g.*daily objects like a paper bag). When multiple sources are present, it can reconstruct the sounds separately with respect to *distance* which could not be achieved by classical beamforming in microphone arrays, while being immune to motion interference. *RadioMic* can also sense sound through walls and even soundproof materials as RF signals have different propagation characteristics than sound. Potentially, *RadioMic*, located in an insulated room (or in a room with active noise cancellation [123]), can be used to monitor and detect acoustic events *outside* the room, offering both sound proofness and awareness at the same time.

To enable these applications, *RadioMic* employs a novel metric to detect the sound events in the environment, while rejecting the non-acoustic interference. Robust detection algorithm

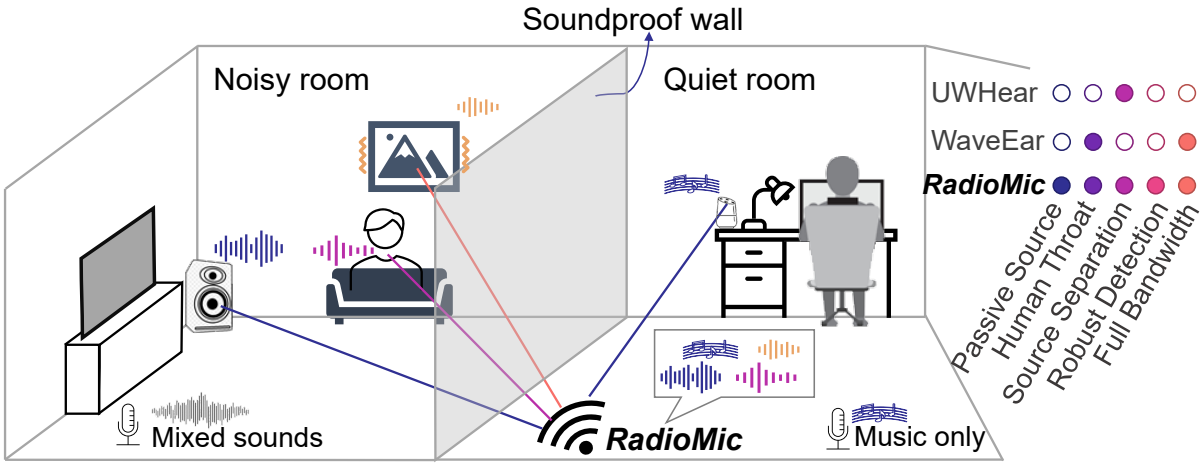


Figure 3.1: An illustrative scenario of *RadioMic*

also enables localization. Furthermore, *RadioMic* also employs a *radio acoustics neural network* to solve the extremely ill-posed high-frequency reconstruction problem, which leverages massive online audio datasets and requires minimal RF data for training.

In summary, this chapter introduces the following:

- Design of *RadioMic*, an RF-based sound sensing system that separates multiple sounds and operates through the walls. To the best of our knowledge, *RadioMic* is the first RF system that can recover sound from passive objects.

- A radio acoustics neural network with a synthetic training method that requires minimal data collection effort, to enhance the sensed sound by expanding the recoverable frequencies and denoising.

- Implementation of *RadioMic* on low-cost COTS hardware and demonstration of multiple attractive applications.

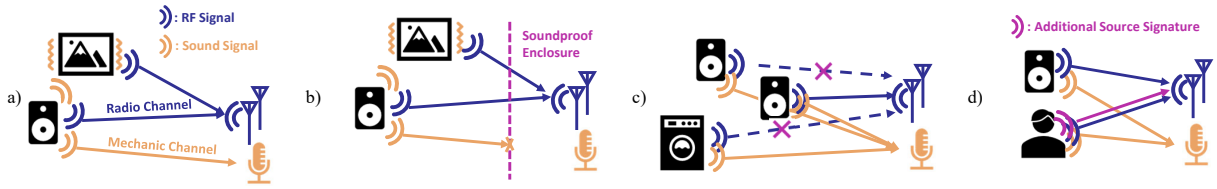


Figure 3.2: Different use cases of *RadioMic*. a) Sensing sound from active/passive sources, b) Sensing through soundproof materials, c) Separating sound of multiple sources, d) Sound liveness detection

3.2 Potential Applications

As in Fig. 3.2, *RadioMic* could benefit various applications, including many that have not been easily achieved before. By overcoming the limitations of today's microphone, *RadioMic* can enhance the performance of popular smart speakers in noisy environments. Collecting spatially-separated audio helps to better understand acoustic events of human activities, appliance functions, machine states, etc. Sound sensing through soundproof materials will provide awareness of outside contexts while preserving the quiet space, which would be useful, for example, to monitor kid activities while working from home in a closed room. Detecting the liveness of a sound source can protect voice control systems from being attacked by inaudible voice [128] or replayed audio [129].

On the other hand, *RadioMic* can be integrated with other existing wireless sensing applications. For example, Soli-based sleep monitoring [17] currently employs a microphone to detect coughs and snores, which may pose privacy concerns yet is no longer needed with *RadioMic*. While remarkable progress has been achieved in RF-based imaging [28, 130, 131], *RadioMic* could offer a channel of the accompanying audio. Some of these applications will be introduced and investigated in the subsequent chapters.

3.3 *RadioMic* Design

RadioMic consists of four main modules. It first extracts CIR from raw RF signals in Section 3.3.1. From there, it detects sound vibration (Section 3.3.2) and recovers the sound Lastly, it feeds the recovered sound into a neural network for enhancement (Section 3.3.3).

3.3.1 Raw Signal Conversion

Our implementation mainly uses a COTS FMCW mmWave radar, although *RadioMic* can work with other mmWave radios that report high-resolution CIR, such as an impulse radar. Prior to explaining how *RadioMic* recovers sound, we provide preliminaries to extract CIR from a linear FMCW radar for a comprehensive explanation. CIR on impulse radar has also been exploited [2, 14, 131]. *RadioMic* relies on the preliminaries explained previously in Chapter 2. To detect and localize sound, *RadioMic* extracts spectrogram representation from each distance bin, as introdeced in Chapter 2 using (2.4).

3.3.2 Sound Detection & Localization

As any range bin can have sound vibration, it is critical to have a robust detection module that can label both *range bins* and *time indices* effectively. Standard methods in the literature, such as constant false alarm rate (CFAR) or Herfindahl–Hirschman Index (HHI) [2] are not robust, as we envision a system to be triggered *only* by sound vibration but not arbitrary motion.

In *RadioMic*, we leverage the physical properties of sound vibration to design a new approach. Mainly, *RadioMic* relies on the fact that a vibration signal creates both *positive* and *negative* Doppler shifts, as it entails consequent displacement in both directions. This is expected

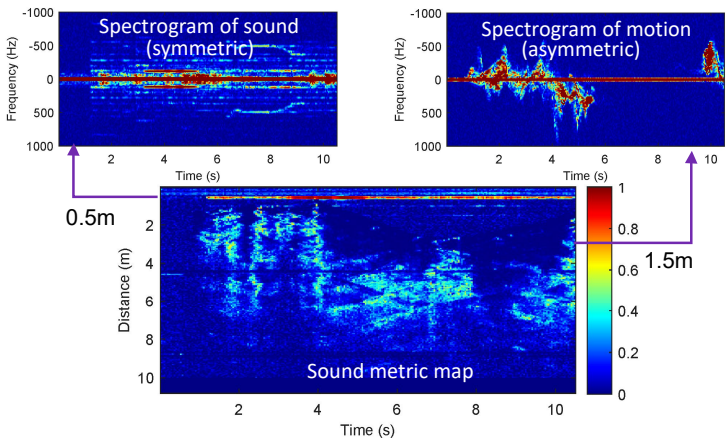


Figure 3.3: Sound metric. An aluminum foil is placed at 0.5m. Music starts playing around 1.5s, while random motion occurs at distances 1∼3m for 10s. *Spectrograms* at distance (a) 0.5m and (b) 1.5m, and (c) resulting *sound metric map*

to result in *symmetric spectrograms*, as also noted by other work [3, 42]. *RadioMic* exploits this observation with a novel metric for robust sound detection.

To define a sound metric, let $G^+(f, r, k)$ denote the magnitude of the positive frequencies of range-Doppler spectrogram $G(f, r, k)$, *i.e.*, $G^+(f, r, k) = |G(f, r, k)|$ for $f \in (0, N_s/2)$. We first subtract the noise floor from both G^+ and G^- and denote the resulting matrices with $\hat{G}^+$ and $\hat{G}^-$ to remove the effect of background. Then, instead of using standard cosine distance, we change the definition to enforce similarity of the amplitude in $\hat{G}^+$ and $\hat{G}^-$:

$$m(r, k) = \frac{\sum_f |\hat{G}^+(f, r, k)\hat{G}^-(f, r, k)|^2}{\max\left(\sum_f |\hat{G}^+(f, r, k)|^2, \sum_f |\hat{G}^-(f, r, k)|^2\right)}. \tag{3.1}$$

RadioMic calculates the *sound metric* as in (3.1) for each range bin r, and for each time-frame k, resulting in a *sound metric map* as illustrated in Fig. 3.3c, music sound (Fig.3.3a) results in high values of the sound metric, whereas arbitrary motion (Fig.3.3b) is suppressed significantly, due to asymmetry in the Doppler signature and power mismatches. This illustrates the responsiveness of sound metric to vibration while keeping comparatively lower values for

random motion.

Based on the sound metric, two different thresholding methods could be implemented, i) Static, hard-coded threshold (denoted by T), ii) Dynamic threshold based on outliers (denoted by O).

After detecting and localizing the source, we can extract the sound signal by using the method presented in Chapter 2. To further improve the quality of the reconstructed sound, *RadioMic* uses various diversity combining methods, including receiver and multipath diversity, which will be denoted by (D) and (R).

3.3.3 Sound Enhancement via Deep Learning

Even though the aforementioned processes reduce multiple noise sources, and optimally create a sound signal from radio signals, there are fundamental limitations. Our results in Chapter 2 indicate that frequencies beyond 2 kHz are attenuated fully in the recovered sound, as the channel h in (2.5) removes useful information in those bands. This creates a significant problem, as the articulation index, a measure of the amount of intelligible sound, is less than 50% for 2kHz band-limited speech [132]. To explain why this happens, we return back to our modeling of the signal. Namely, the output signal $\hat{x}(t)$ is a noisy copy of $x(t)$, which could be written as:

$$\hat{x}(t) \approx x(t) + \hat{w}(t) = h \star a(t) + \hat{w}(t), \tag{3.2}$$

from (2.5). As can be seen, what we observe is the output of the air pressure-to-object vibration channel (or mechanical response of a speaker), as also observed by [1, 44, 66, 70].

In order to recover $a(t)$ fully, one needs to invert the effect of h. However, classical signal

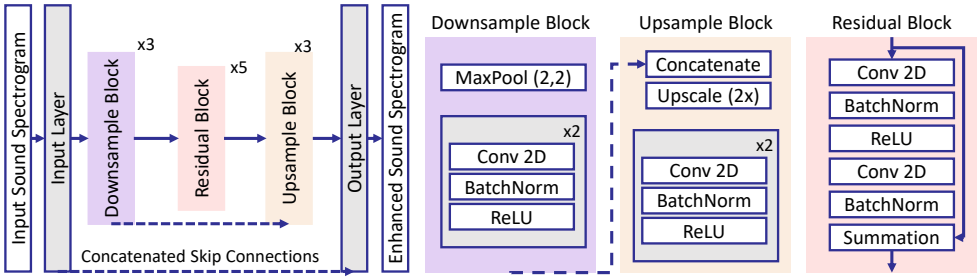


Figure 3.4: RANet Structure

processing techniques like spectral subtraction or equalization cannot recover the entire band, as the noise levels at high frequencies are extremely high. To overcome these issues, we build an autoencoder based neural network model, named as *radio acoustics networks* (RANet). Although the formulation of the problem is similar to those in sound enhancement and denoising domain (*e.g.* [133, 134]), theoretical limitations are stricter in *RadioMic*, as there is severer noise, and stronger band-limit constraints on the recovered speech (expanding 2 kHz to 4 kHz), in addition to the need for solving both problems together rather than studying them separately. Fig. 3.4 portrays the structure of RANet, with the entire processing flow of data augmentation, training, and evaluation illustrated in Fig. 3.5.

1) RANet Structure: RANet consists of downsampling, residual, and upsampling blocks, which are connected sequentially, along with some residual and skip connections. On a high level, the encoding layers (downsampling blocks) are used to estimate a latent representation of the input spectrograms (*e.g.* similar to images); and decoding layers (upsampling blocks) are expected to reconstruct high-fidelity sound. Residual layers in the middle are added to capture more temporal and spatial dependencies by increasing the receptive field of the convolutional layers, and to improve model complexity.

2) Training RANet without Massive RF Data: A successful training process for a relatively

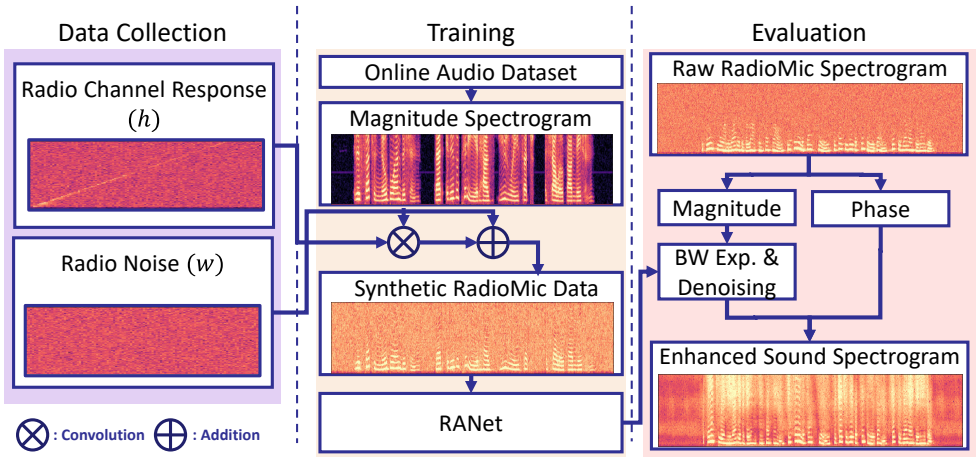


Figure 3.5: Working process of RANet in *RadioMic*

deep neural network as proposed here requires extensive data collection. However, collecting massive RF data is costly, which is a practical limitation of many learning-based sensing systems. On the other hand, there have been growing, massive audio datasets becoming available online. In *RadioMic*, instead of going through an extensive data collection procedure like [48], we exploit the proposed radio acoustics model and translate massive open-source datasets to synthetically simulated radio sound for training. Two parameters are particularly needed to imitate radio sound with an audio dataset, *i.e.*, the channel h and noise w as in (3.2). We use multiple estimates for these parameters to cover different scenarios and artificially create radar sound at different noise levels and for various frequency responses, thus allowing us to train RANet efficiently with limited data collection overhead.

3) Generating Sound from RANet: Using the trained model, RANet uses raw radar sound as input and extracts magnitude spectrograms that will be used for denoising and bandwidth expansion. Output magnitude spectrograms of RANet are combined with the phase of the input spectrograms, as usually done in similar work [135, 136] and the time-domain waveform of the speech is constructed.

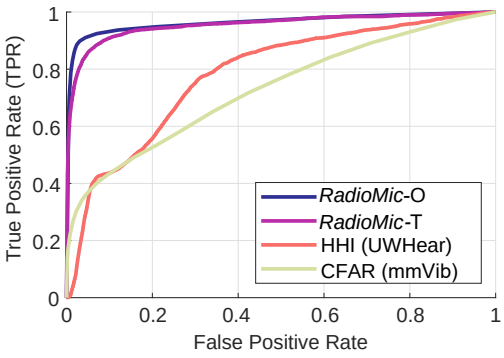


Figure 3.6: ROC curve of sound detection

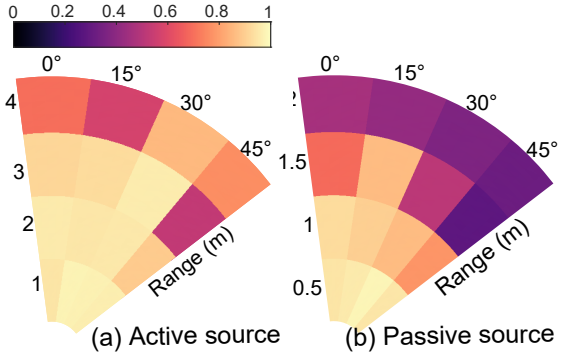


Figure 3.7: Detection coverage of *RadioMic*

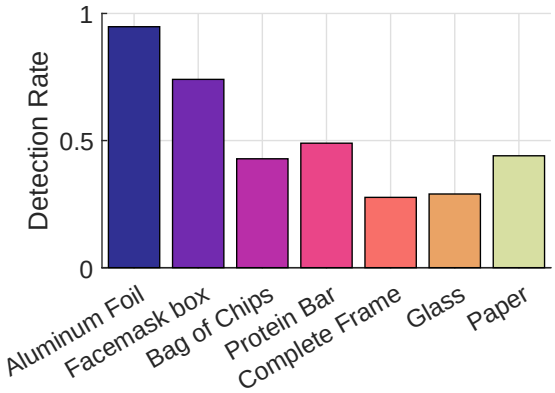


Figure 3.8: Detection with different daily materials

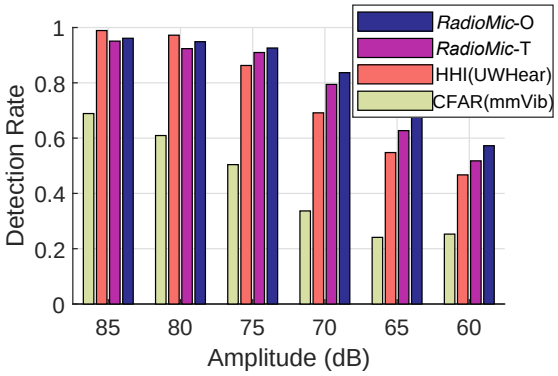


Figure 3.9: Detection at different sound levels

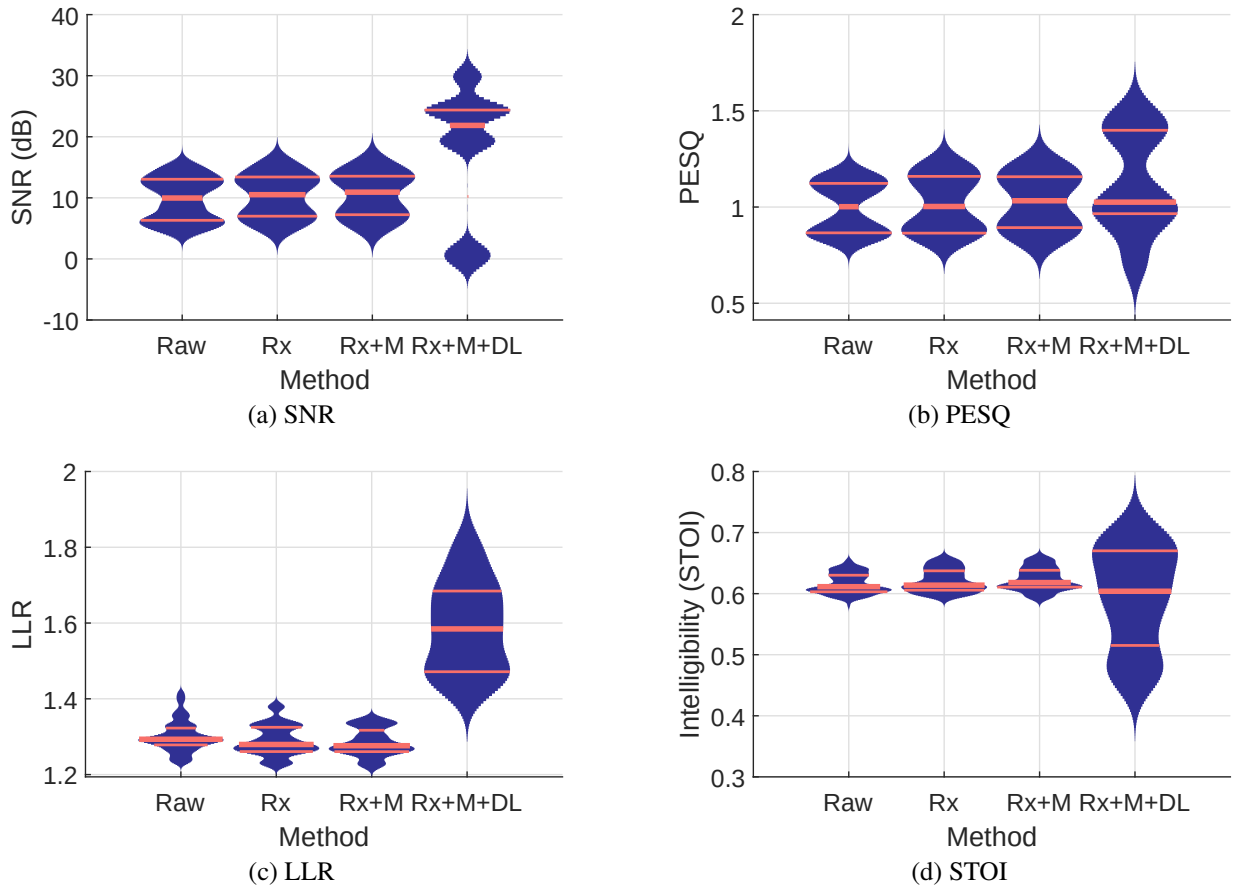
3.4 Experiments and Evaluation

We benchmark different modules of *RadioMic* in multiple places, such as office space, home, and an acoustically insulated chamber. To evaluate the performance of *RadioMic*, we utilize metrics signal-to-noise ratio (SNR), STOI [126], perceptual quality (PESQ) [125] and LLR [127]. Furthermore, we also visualize spectrograms to better illustrate the outputs of *RadioMic*.

3.4.1 Detection and Localization

Our data collection for detection analysis includes random motions, such as standing up and sitting down repeatedly, walking, running, and rotating in place, as well as static reflectors in the environment and human bodies in front of the radar. On the other hand, we also collect data using multiple sound and music files with active and passive sources. More importantly, we have also collected motion and sound data simultaneously to see if *RadioMic* can reject these interferences successfully.

To illustrate the gains coming from the proposed sound metric, we implement and compare with existing methods: 1) HHI (UWHear [2]): UWHear uses HHI, which requires some training to select an appropriate threshold. 2) CFAR (mmVib [3]): To imitate the same approach and provide a reasonable comparison, we apply the classical CFAR detection rule at various threshold levels and remove the detections around DC to have a fairer comparison. Additionally, we also compare hard thresholding (*RadioMic*-T) with the outlier-based detector (*RadioMic*-O), in our system. We provide the receiver-operating characteristics (ROC) curve for all methods in Fig. 3.6. As can be seen, while *RadioMic*-T is slightly worse than *RadioMic*-O, the other methods fail



(a) SNR (b) PESQ (c) LLR (d) STOI

Figure 3.10: Overall performance of *RadioMic* with gains from multiple components. Rx: receiver combining; Rx+M: receiver and multipath combining; Rx+M+DL: the end results.

to distinguish random motion from the vibration robustly, which prevents them from practical applications as there would be arbitrary motion in the environment.

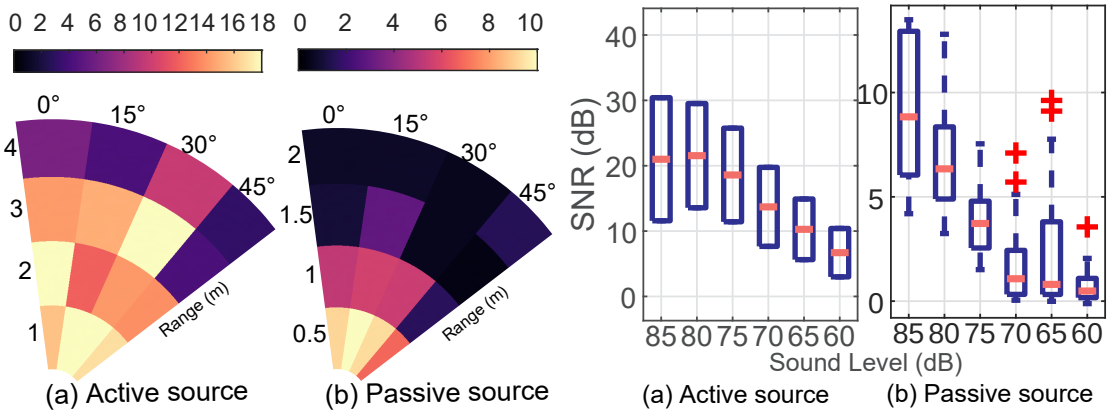
Materials of Passive Sources: To further evaluate the detection performance of *RadioMic* with passive materials, we conduct experiments with daily materials, such as picture frames, paper bags, or bags of chips. As shown in Fig. 3.8, many different materials enable sound detection using *RadioMic*. And even at a lower rate, some sound signal is detected for particular instances as the evaluation is done with frames with 40 ms duration, with outputs a decision every 10ms. We also present the results with longer windows of 500 ms for a more practical application. As the lower frequencies are more prominent in the outputs of *RadioMic*, it is safe to assume that

basic features (*e.g.* pitch) could be extracted from many materials in the environment.

Operational Range: Since previous art on the material study reveals that it is difficult to capture much sound content using a variety of paper and glass objects [1, 44, 66], we shift our focus to operational range using a relatively better object, aluminum foil. We investigate the detection performance of *RadioMic* at different distances azimuth angles (with respect to the radar) using an active source (*a pair of speakers*) and a passive source (*aluminum foil of size* 4×6 inches). We use 5 different sound files for each location, three of which are music files and two are human speech. As shown in Fig. 3.7, *RadioMic* can robustly detect sound up to 4m in an active case with 91% mean accuracy, and up to $2m$ in the passive source case with %70 accuracy, both with a field of view of 90°. Passive source performance is expected to be lower, as the vibration is much weaker.

3.4.2 Sound Reconstruction Performance

Overall Performance: With gains from diversity combining and deep learning, we provide the overall performance of *RadioMic* in Fig. 3.10. We investigate the effect of each component on a dataset using a *passive source*. Overall, each of the additional diversity combining schemes improves the performance with respect to all metrics. At the same time, RANet reduces the total noise levels significantly (Fig. 3.10a) and increases PESQ (Fig. 3.10b). However, as in Fig. 3.10c, RANet yields a worse value with LLR, which is due to the channel inversion operation of h applied on the radar signal. While an optimal channel recovery operation is demanded, RANet is trained on multiple channel responses and only approximates to h. Consequently, the channel inversion applied by RANet is expected to be sub-optimal. Lastly, STOI metric (Fig. 3.10d)



(a) Locations (b) Amplitudes

Figure 3.11: Recovered sound SNR (a) at different locations and (b) with different sound amplitudes.

shows a higher variation, which is due to high levels of noise in the sample audio files in the input. In the case of large noise, we have observed that RANet learns to combat the effect of noise w, instead of inverting h, and outputs mostly empty signals, which could also be observed by the distribution around 0 dB in Fig. 3.10a. When there is enough signal content, RANet improves the intelligibility further.

Distances and Source Amplitudes: To investigate sound recovery from varying locations and angles, we provide two heatmaps in Fig. 3.11a to show the raw SNR output for active and passive sources. Similar to sound detection in Fig. 3.7, nearby locations have higher SNR, allowing better sound recovery, and the dependency with respect to the angle is rather weak. Increasing distance reduces the vibration SNR strongly, (*e.g.*, from 20 dB at 1m to 14 dB at 2m for an active source) possibly due to the large beamwidth of our radar device and high propagation loss. Generally, the SNR decreases with respect to decreasing sound levels. And at similar sound levels, a passive source, aluminum foil, can lose up to 10 dB compared to an active source. In addition, *RadioMic* retains a better SNR with decreasing sound levels than increasing distance (Fig. 3.11a), which

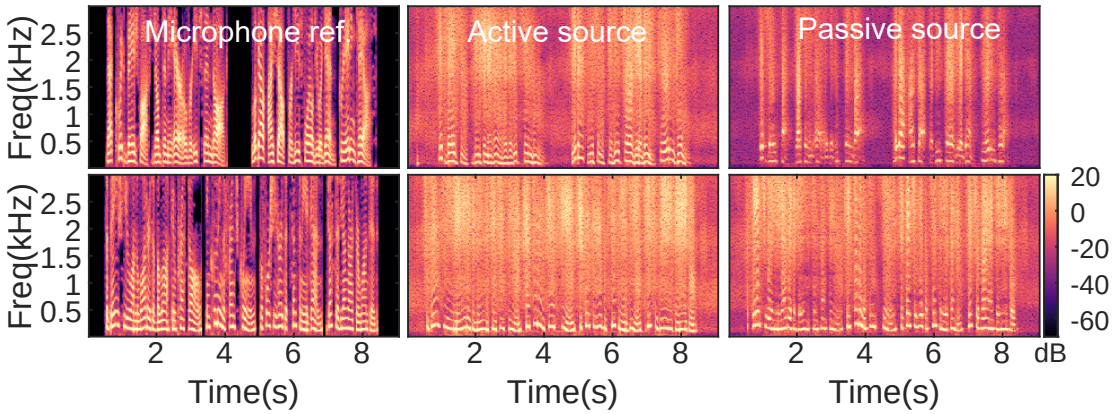


Figure 3.12: Spectrogram comparison of *RadioMic* outputs and a microphone. Two rows correspond to the synthesized speech of two different sentences. Passive source is a small aluminum foil, whereas active is a loudspeaker.

indicates that the limiting factor for large distances is not the propagation loss, but the reflection loss, due to relatively smaller surface areas. Hence, with more directional beams (*e.g.* transmit beamforming, or directional antennas), the effective range of the *RadioMic* could be improved, as low sound amplitudes also look promising for some recovery. In Fig. 3.12, we also provide exemplary spectrograms for active and passive objects.

Active vs. Passive Comparison: In order to show potential differences between the nature of active and passive sources, and present the results in a more perceivable way, we provide six spectrograms in Fig. 3.12, which are extracted by using two different synthesized audio files. In this setting, the passive source (aluminum foil) is placed at 0.5m away and the active source is located at 1m. As shown, the active source (speaker diaphragm) has more content in the lower frequency bands, whereas passive sound results in more high-frequency content, due to the aggressive channel compensation operation on h. More detailed comparisons are provided in Table 3.1.

LOS vs. NLOS Comparison: We further validate *RadioMic* in NLOS operations. To that end, in addition to our office area, we conduct experiments in an insulated chamber (Fig. 3.13c),

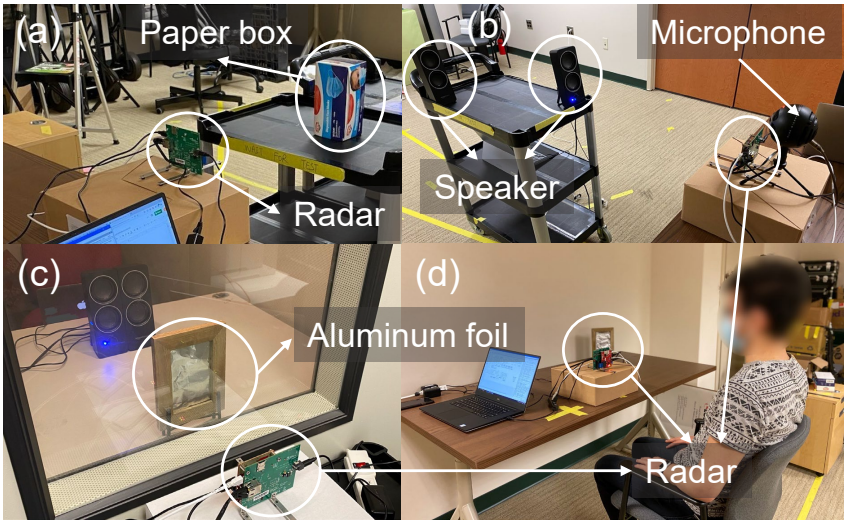


Figure 3.13: Example setups. (a) Passive source; (b) Multiple speakers; (c) Insulated chamber; (d) Sensing from throat.

Table 3.1: Active vs. Passive Source Comparison

Setup	SNR	PESQ	LLR	STOI
LOS, Active	24.7	0.84	1.61	0.55
LOS, Passive	10.4	1.20	1.57	0.61
NLOS, Active	29.4	1.12	1.52	0.58
NLOS, Passive	8.8	1.36	1.57	0.64

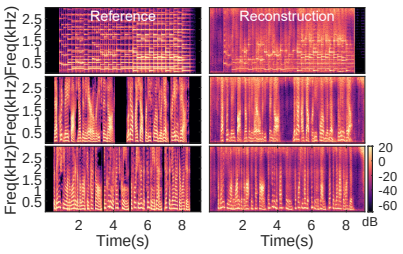


Figure 3.14: Through-wall spectrograms. Left: microphone reference; Right: reconstructed results. The top row also includes a music file.

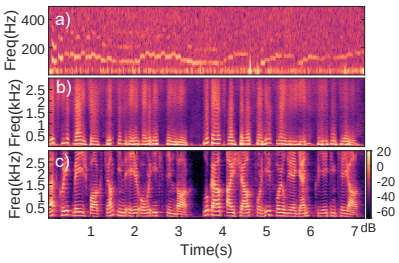


Figure 3.15: Recovery from the throat. *RadioMic* spectrogram of a) humming a song around 60 dB, and b) speaking, c) Microphone spectrogram for case b).

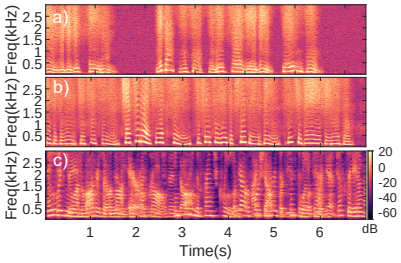


Figure 3.16: Multiple source separation. Spectrograms of *RadioMic* for a) source #1 and b) source #2, c) Microphone spectrogram with mixed sound.

which has a double glass layer on its side. This scenario is representative of expanding the range of an IoT system to outside rooms from a quiet environment. In this particular scenario, we test both the passive source (*e.g.* aluminum foil), and the active source (*e.g.* speaker). As additional layers would attenuate the RF reflection signals further, we test NLOS setup at slightly closer distances, with the active speaker at 80 cm and the passive source at 35 cm away. Detailed results are in Table 3.1, with visual results in Fig. 3.14. As seen, insulation layers do not affect *RadioMic* much, and LOS and NLOS settings perform quite similarly. Some metrics even show improvement in the NLOS case due to shorter distances. These results are valid for double glass, which does not attenuate RF signals much, but are not applicable when a metallic sheet is used around the insulation, as metals can attenuate the RF signals significantly.

Comparison with VisualMic [1]: Here, we provide performance metrics of *RadioMic*, using the reference audio files provided by the authors of [1]. In Table 3.2, we provide the metrics from resulting files of *RadioMic*, along with those from VisualMic. This is not a direct comparison, as the two systems are using completely different modalities and hardware (VisualMic needs expensive high-speed cameras). Instead, we report the performance figures reported by the authors.

Table 3.2: Comparison using files in VisualMic [1] with *RadioMic*

Sequence	Method	SSNR	LLR	STOI
Female - fadg0,sa1	*RadioMic*	17.94	1.36	0.48
	VisualMic	24.5	1.47	0.72
Female spk. - fadg0,sa2	*RadioMic*	30.05	1.15	0.49
	VisualMic	28.7	1.37	0.65
Male spk. - mccs0,sa1	*RadioMic*	33.38	1.67	0.48
	VisualMic	20.4	1.31	0.59
Male spk. - mccs0,sa2	*RadioMic*	9.46	1.94	0.42
	VisualMic	23.2	1.55	0.67
Male spk. - mabw0,sa1	*RadioMic*	27.55	1.18	0.58
	VisualMic	23.3	1.68	0.77
Male spk. - mabw0,sa2	*RadioMic*	24.63	1.11	0.58
	VisualMic	25.5	1.81	0.72

Although we cannot make a direct comparison, the results indicate that, similar sound reconstruction performance could be achieved by a $15 mmWave radio device, compared to an expensive high speed camera, which starts from $5,000 [137], and can be as expensive as $100,000 [138].

Sound Recovery from Human Throat: Lastly, we show how *RadioMic* can also capture vocal folds vibration from the human throat as another active source. We start with humming in front of the speaker at a quiet 60 dB level, when the person sits 1m away from the device in the direct LOS and show the results in Fig. 3.15a. After this observation, we collect multiple recordings from a user, where the setup is given in Fig. 3.13d. The person sits 1m away from the device in the given setting and reads multiple sentences. In Fig. 3.15b and 3.15c, we provide the *RadioMic* and microphone spectrograms. Although *RadioMic* is not trained with a frequency response h from the human throat, it can still capture useful signal content, such as the first few harmonics. On the other hand, we noticed that the intelligibility of such speech is rather low, compared to other sources (*e.g.*, daily objects). Prior work [48] focuses on extracting sound from the human throat, and with extensive RF data collection, they have shown the feasibility of sound recovery

from the throat. We believe that *RadioMic* could be used with the same mode of operation (*i.e.*, sound recovery from human throat) as well with collecting and training with massive RF data from human throat. On the other hand, we would like to point out the fundamental differences between sound recovery from a human throat and objects:

- **Frequency response:** Air-vibration-to-object-displacement channel could be considered as a linear time-invariant channel that does not depend on the input signal. On the other hand, according to the source-filter model [122], source (vocal folds) and filter (vocal tract) can be changed independently. In other words, different sound signals (*e.g.*, $\backslash a \backslash$ and $\backslash e \backslash$) can share the same source (vocal folds vibration), but different filters (shape of the vocal tract). The filter affects the entire frequency range, and therefore, the speech-to-radar-captured-vocal-folds-vibration channel could be considered as a signal-dependent channel, and capturing the vocal folds using radar would not be sufficient for recovering the speech completely.

- **Throat motion and data simulation:** Our training method simulates the object vibration by capturing frequency responses from objects through playing a frequency sweep. It is not possible to simulate the throat signals with high accuracy in the same way, as the simulation loses motion information from the throat. The motion cannot be modeled as simple as frequency response and would require extensive modeling and/or data collection. With the help of extensive data collection and this side information, *RadioMic* should be able to reconstruct the sound fully, as done in WaveEar [48].

Since the feasibility of such a mode of operation is shown, we leave these improvements for the future and switch our focus to another application of sound liveness detection of human

subjects in the next section.

3.5 Case Studies

In this section, we show the characteristics of *RadioMic* for multiple source separation in Section 3.5.1 and then extend it to classify sound sources in Section 3.5.2. Later on, we discuss potential privacy issues and countermeasures in Section 3.5.3.

3.5.1 Multiple Source Separation

The separation of multiple sound sources would enable multi-person sensing, or improved robustness against interfering noise sources in challenging conditions. In recent years, deep learning based methods have shown remarkable progress on separating speech in multitalker scenarios, or denoising a single speaker sound. These methods usually constrain the problem into certain tasks, and universal sound separation is still an open problem [139]. Having a microphone array solves sound source separation problems when the interferers are spatially separated from the source, yet many challenges still exist when the environment is challenging (*e.g.* noisy or reverberant) [75], or when the interferers are close to the source in the azimuth domain. Under this setting, *RadioMic* brings a new direction by sensing the sound at the source, and can potentially separate the sources not only in the azimuth domain, but also in the distance domain; and it has no constraints on the type of sound.

In order to illustrate the feasibility of sound separation using *RadioMic* with respect to *distance* using *RadioMic*, we play two different speech files simultaneously from the left and right channels of the stereo speakers. As shown in Fig. 3.13b, we place the right speaker at 0.75m, and

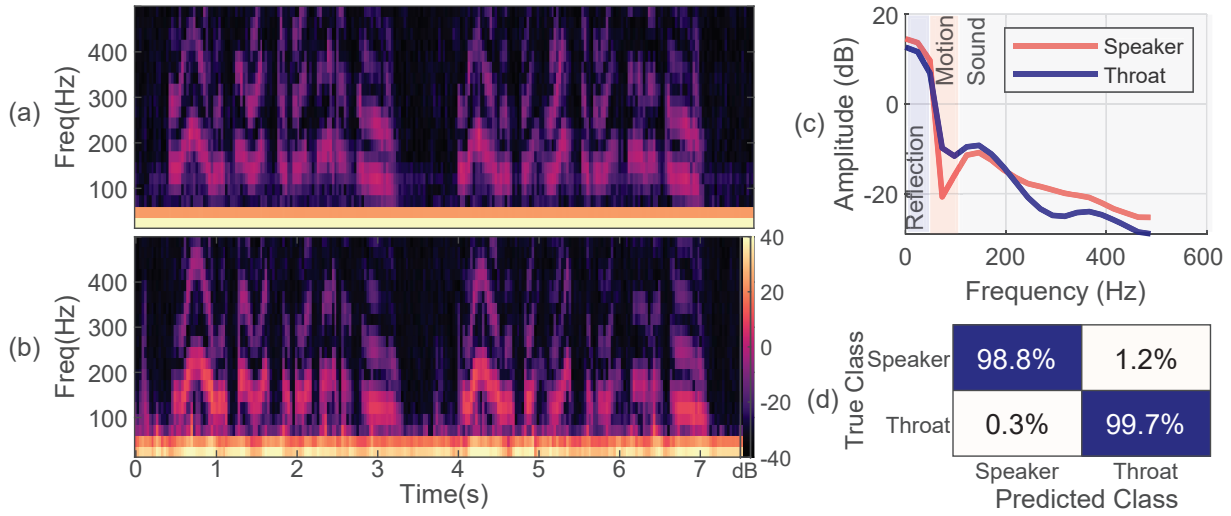


Figure 3.17: (a) Speaker spectrogram, (b) throat spectrogram, (c) Power delay profile extracted from (a,b), (d) confusion matrix for classification.

the left speaker at 1.25m. We provide the results in Fig. 3.16, which include two spectrograms extracted by *RadioMic*, along with a microphone spectrogram. As seen, the microphone captures a mixture of multiple sources and is prone to significant interference. In contrast, *RadioMic* signals show much higher fidelity, and two person's speech can be separated from each other well. In this case study, we do not pursue the separation performance with respect to the azimuth domain, as a similar technique (beamforming) can be applied with microphone arrays in good environmental conditions, and a direct comparison of *RadioMic* with a microphone array would require an extensive evaluation. Previous work UWHear [2] specifically focuses on the problem of sound separation and demonstrates good performance using UWB radar. *RadioMic* excels in achieving more features in one system, in addition to the source separation capability. And we believe there is a great potential to pursue higher fidelity by using *RadioMic in tandem with a single microphone*, and to understand physical limits in terms of azimuth and distance separation, which we investigate in Chapter 4.

3.5.2 Sound Liveness Detection

As another application, we investigate the feasibility of sound source classification. As *RadioMic* senses at *source* of the sound, it captures the additional physical characteristics of the *sound generation mechanism* simultaneously. Starting with this observation, we investigate the question: *Is it possible to differentiate the source of a sound between a human and an inanimate source like a speaker?* This is a critical application as it is well-known that today's microphones all suffer from inaudible attacks [128] and replay attacks [123] due to hardware defects.

Our results show that *RadioMic* can enable sound liveness detection with unprecedented response times. In our experiment, we ask a user to sit around 1m away from the device and speak at a normal speaking volume (around 70dB at 0.5m away) and recite five different sentences, in two different languages, and we record the speech using a condenser microphone Afterward, we play the same sound through speakers at the same distance, at a similar sound level, and capture *RadioMic* output. The data collection setting is similar to that in Fig. 3.13, but with increased distance and an additional microphone for sound recording.

We first provide the comparison of two raw spectrograms (*i.e.*, without filtering and projection operation explained in Eq. (2.8)) in Fig.3.17(a,b), and the average over time in Fig. 3.17(c). From the figures, we make three observations: 1) As illustrated by *reflection* band in Fig. 3.17(c), the human throat shows a weaker amplitude around DC component. This is because the reflection coefficients of speakers and human throat vary significantly, a phenomenon utilized for material sensing [14]. 2) Due to minute body motions and the movement of the vocal tract, the reflected signal energy from the human throat varies more over time and has stronger sidelobes, which could be seen in the frequency band labeled as *motion* in Fig. 3.17(c). 3) Due to skin layer

between vocal cords and the radar, the human throat applies stronger low-pass filtering on the vibration compared to speakers, as labeled as *sound* in Fig. 3.17(c), which relates to the frequencies of interest for sound.

Then to enable *RadioMic* for liveness detection, we implement a basic classifier based on these observations. We propose to use the ratio of the energy in motion affected bands (35-60 Hz) over the entire radar spectrogram as an indicator for liveness. As shown in Fig. 3.17(d), *RadioMic* can classify the sources with 95% accuracy with only **40 ms** of data, which increases to 99.2% by increasing to $320ms$. We believe *RadioMic* promises a valuable application here as it can sense the sound and classify the source *at the same time*, and we plan to investigate it thoroughly in the future.

3.5.3 Privacy Considerations

As discussed and illustrated previously, *RadioMic* can sense sound through certain soundproof enclosures, such as glass, due to different propagation characteristics of RF waves, compared to acoustic signals. Therefore, *RadioMic* can raise privacy concerns, as it enables sensing in challenging conditions. An attacker can potentially use *RadioMic* to steal sensitive data from a distance.

To that end, we explore the usefulness of shielding to protect the privacy and prevent data leakage. We experiment with a steel mesh to show that, a metallic enclosure can reduce the capacity of *RadioMic* significantly. For our experiments, we use a paper enclosure, which does not block RF signals for reference, and compare it with a steel enclosure, wrapped outside the original paper. We place the speakers at a close distance of 50cm, play a single-tone sound file at

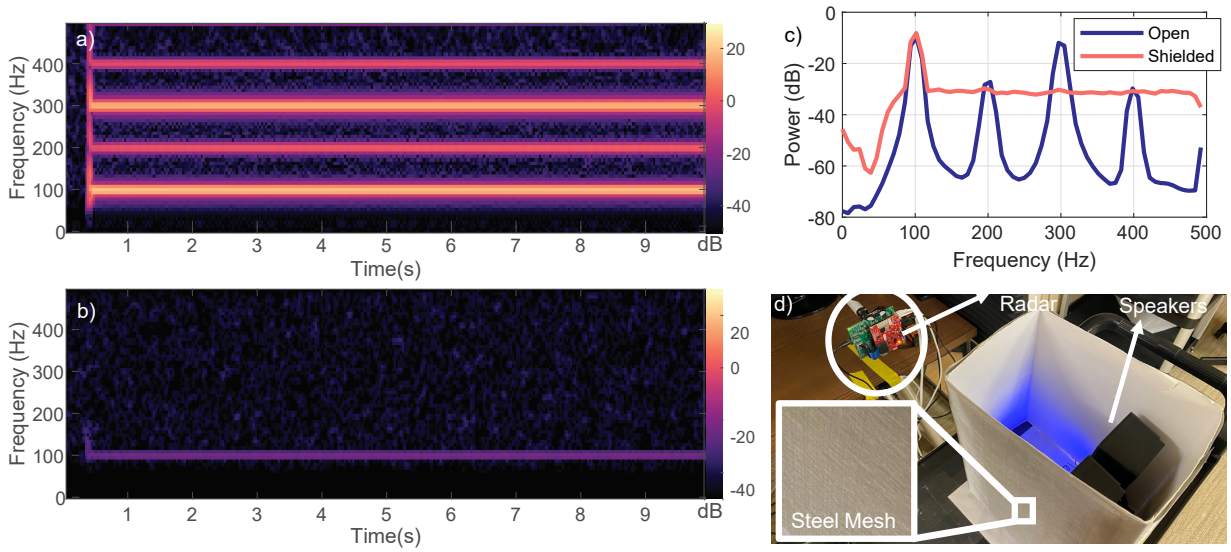


Figure 3.18: (a) Spectrogram of reconstructed sound (open case), (b) Spectrogram of reconstructed sound when insulation is present (shielded case), (c) Power delay profile extracted from (a,b) after normalization, (d) Experimental setting.

100 Hz, and compare the recovered signal in two scenarios. We add tension on the steel mesh to reduce its vibration due to sound (*i.e.*passive object recovery). In this experiment, we manually enter the actual location of speakers, instead of using the localization module of *RadioMic*, as the reflections are extremely weak.

As shown in Fig. 3.18(a) and Fig. 3.18(b), a basic steel mesh reduces the recovered signals considerably, where high frequencies are completely removed, and the first harmonic at 100Hz is reduced substantially. Furthermore, after normalizing the amplitudes of signals, we average the two plots in Fig. 3.18(c), and observe that the noise floor is increased roughly $30dB$ for (equivalent to signal amplitude reducing the same amount), and higher order harmonics cannot be observed. We provide the experimental setting in Fig. 3.18(d), and conclude that a metallic mesh can be sufficient to *reinsulate* the private spaces against *RadioMic*.

3.6 Discussion

RadioMic enables a new and comprehensive capability of sensing the environment by recovering sound from the radio signals, which opens new directions and promises exciting applications. As a pioneering effort, however, there leaves room for improvements in the following areas.

Sound detection: *RadioMic* uses a novel metric to detect the presence of sound on object surfaces. Although sound metric can successfully capture sound-induced vibration, and reject random motion in the environment, it can also capture other sources of vibration, such as a fan running in the background. Regular/constant vibration sources can be eliminated by background subtraction on sound metric; however, irregular vibration would require further characterization and classification. This could be mitigated by a classification module at the end of the pipeline of *RadioMic*, yet we leave this to future work.

Source motion: *RadioMic* is designed to detect and localize stationary and near-stationary sources, and to reject moving objects and their interference. Although this is a practical approach for speakers and daily objects in the environment, moving sound sources cannot be detected, due to the design of sound metric. In a similar task of capturing heart-rate signals from moving human bodies, methods such as empirical mode decomposition [19] are used to reduce the effect of motion. We think that sound sensing is a more challenging problem compared to heart rate, as the amount of displacement is shorter ($60 - 180\mu m$ for vocal folds vibration [140], compared to $200 - 500\mu m$ for heart-rate induced displacement [141]). Furthermore, the speech signal is less constrained compared to regular heart beats, and we believe there is plenty of room for improvement to reject the body motion, and sense sound-related vibration from moving bodies.

Deep learning based sound enhancement and denoising: *RadioMic* uses a basic deep learning model RANet to denoise and enhance bandlimited sound. Although these two problems are extremely popular in their corresponding domains, to the best of our knowledge, there is minimal work investigating both problems *at the same time*, especially for extremely narrowband speech (¡2kHz). Given the increasing interest in constructing sound from ambient objects using different modalities recently ([1,66,68,70]), where the constructed raw signals are extremely bandlimited, we believe more sophisticated deep learning systems could be developed, and all these modalities can benefit from the improvements. Our training procedure in 3.3.3 would enable using public datasets to compare different structures and further improvements could be obtained in terms of the performance.

Performance: In this Chapter, we utilize a COTS device to reconstruct the sound from various sources. In our experiments, we realize that our effective range is limited to short distances (*e.g.*, ¡4m), mainly due to relatively wide beamwidth of our device ($> 30°$). Other works focusing on extracting throat vibration either use highly directional antennas (even up to $1°$) beamwidth [46], very close distance (less than 40 cm) in [142], many more antennas (*e.g.*, 16×16) [48]. We believe that more advanced hardware and sophisticated beamforming could underpin the better performance of *RadioMic*. Better hardware would enable successful sound recovery for low amplitude sound and longer distances, and would make the distance limits for applications in Section 3.5 more practical.

Multimodal systems: We believe *RadioMic* and microphones are complementary. Instead of leveraging radar information solely to reconstruct sound, the side information from *RadioMic* could be used in tandem with a microphone to achieve better performance of sound separation and noise mitigation than the microphone alone. Similar to audio-visual sound enhancement [93],

an *audioradio* based system could achieve superior performance compared to a microphone. Furthermore, side information from *RadioMic* can enable sophisticated security systems with sound liveness detection against side-channel attacks. Exploring *RadioMic* with mmWave imaging [131] and other sensing [17] is also an exciting direction. As another relevant direction, in Chapter 4, we pursue a multimodal speech enhancement and separation system that relies on the side channel of mmWave-based sensing, with a different processing pipeline due to practical considerations.

Privacy and security considerations: *RadioMic* is shown to have different characteristics than microphones, due to the radio modality. As discussed in Section 3.5, an attacker can potentially use *RadioMic* to eavesdrop into assumed-to-be-private environments and steal sensitive data. Therefore, more advanced shielding methods are needed in certain applications. On the other hand, using *RadioMic* instead of a microphone brings a different notion of privacy. *RadioMic* only senses the sound in the field-of-view of the device. This can be used to create sound-sensible locations in a room to create private spaces without relying on software (or hardware) to disable a microphone, which would capture sound from any location within a room.

Practical Considerations: In *RadioMic*, we focus on illustrating the feasibility of sound sensing via radio signals and there is room for improvement to make the system more practical, in terms of power consumption and computational complexity. As discussed in RANet, more efficient and compact deep learning systems can be proposed to reduce the computational complexity. Since our focus has not been on building a real-time system, there is a processing delay of 1s, which could be improved. To reduce the computational complexity for band-limited applications, the sampling rate can be reduced, which would also decrease power consumption. Current localization and detection algorithm runs on extracting sound metric for all range bins, which can include

a possible preprocessing (*e.g.*CFAR based filtering) to reduce the complexity further. Last, co-existence of other radio devices is also a practical consideration. Our device runs at 77GHz, which is not used by communication devices. The interference from other mmWave radar devices can be mitigated by careful placement of devices, as these signals have high directionality. In addition, the coexistence problem has been an interesting problem in radar literature [143], and some of the interference-management methods can be used with *RadioMic*.

3.7 Summary

In this chapter, we propose *RadioMic*, an mmWave based-based sound and vibration sensing system that can reconstruct sound from sound sources and passive objects in the environment. *RadioMic* is robust against environmental changes, such as those in lighting, and can operate in dark and NLOS settings, such as through an insulated layer of glass. Using micrometer level vibrations that occur on the object surfaces due to a sound source, *RadioMic* can detect and recover sound, as well as identify sound sources using a novel radio acoustics model and neural network. The flexible design of *RadioMic* enables additional applications, such as sound source localization and classification with unprecedented accuracy, separation of sound sources with respect to *distance*, and potentially increasing the sound sensing range to multiple rooms. Extensive experiments in various settings show that *RadioMic* outperforms existing approaches significantly and enables many new applications.

Chapter 4: *RadioSES*: mmWave Based Audioradio Sound Enhancement and Separation System System

4.1 Introduction

Humans are enormously capable of understanding a noisy speech or separating one speaker from another, we collectively refer to these capabilities as speech enhancement and separation (SES), and is known as the cocktail party problem [144]. SES capability for computers is of great demand for many applications, such as voice commands, live speech recording, etc., yet remains a challenging problem using microphones.

Monaural SES methods achieved remarkable progress in the recent years with the help of deep learning, especially when there is not much background noise [145]. However, fundamental problems still exist in estimating the number of sources in a mixture, associating output sources with the desired speakers (*a.k.a* label permutation problem), and tracing the speakers for long periods of time. Single-channel (monaural) methods require estimation of the number of sources, and a robust source association method. Although these problems can be solved for clean mixtures, by clustering-based methods [90] and permutation invariant training (PIT) [87], their performance can decrease with noisy mixtures. Overall, audio-only approaches suffer from these ill-posed problems inherently.

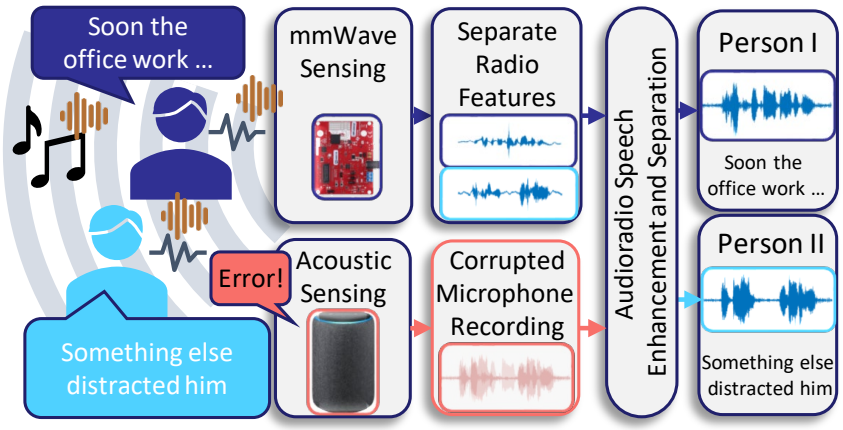


Figure 4.1: *RadioSES* Overview

To overcome the problems and enhance SES, multimodal systems have been introduced to exploit readily available information beyond audio, such as video [93, 94]. Similar to human perception, which also uses lip motion and facial information [146], audiovisual systems are shown to improve SES performance, especially in challenging cases, such as same-speaker mixtures. Same and similar-speaker mixtures are especially difficult for audio-only methods, as the distinction between the two sources is minimal. Additional visual information about the speaker, e.g., videos or even a facial picture of the user [97], or other information, such as voice activity detection [147], or pitch [148] improves the SES performance. However, camera-based methods require good lighting conditions and raise potential privacy concerns.

In this work, we propose to address the SES problem by jointly leveraging millimeter-wave (mmWave) sensing as an orthogonal radio modality. Compared to cameras, radio devices are low-power, can operate in dark, through-wall settings and are less privacy-invasive. The radio reflections from speakers not only can allow separation of multiple speakers but also capture articulatory motions for SES. The reasons to select mmWave radios are two-fold: On the one hand, more and more smart devices now include an mmWave radar and a microphone, such as Google

Soli phone and Nest Hub [16, 17], Amazon Alexa [149] etc. mmWave sensing promises to be more ubiquitous in the future. On the other hand, mmWave sensing has enabled many applications related to motion and vibration, such as heart rate monitoring [37], measuring machinery and object vibration [3, 44], or extracting vocal folds vibration [46]. In particular, it has been used to estimate pitch and detect voice activity [46], reconstruct speech to some extent [45, 48], as well as enhance speech recognition for a single speaker [51]. Yet no existing work has explored utilizing both modalities for *joint* SES tasks.

With this motivation, we develop an *audioradio*[1] speech enhancement and separation system to solve the aforementioned problems and improve the overall performance. Building an audioradio SES system faces multiple challenges. First, in order to solve the number of sources problem, a robust and efficient source detection and tracking method is needed, as the performance of a system can decrease significantly in the event of miss detection. Second, radio signals are usually prone to environmental effects, and their performance can decrease considerably when tested at a new location. Returned signals from the objects are not only affected by vibration, but also from motion, with motion usually being the stronger effect. Third, different from the rich literature in audiovisual deep learning methods, radio modality has not been explored in the context of SES. Designing a suitable and efficient deep learning model for practical applications is non-trivial. Last, deep learning systems require extensive data collection and robust training methods, which is especially challenging for radio signals.

We overcome these challenges in *RadioSES*, the first Audio-**Radio** **S**peech **E**nhancement and **S**eparation system. As illustrated in Fig. 4.1, *RadioSES* can detect, localize, and estimate the

[1] We combine audio and radio words as *audioradio* to refer to a multimodal system consisting of both modalities, similar to the word *audiovisual*.

number of sources in an environment and improve SES performance even in unseen/challenging conditions. To achieve robust detection and localization, we first develop a computationally efficient pipeline of signal processing that can extract the radio features for speakers separately. Then we design an audioradio deep learning framework that takes both audio and radio signals as the inputs and outputs separated and enhanced speeches for each of the speakers. Following recent advances in monaural SES, our deep learning module, called RadioSESNet, utilizes adaptive encoders, instead of relying on classical Short-Term Fourier Transform (STFT) representation. We further introduce a variety of techniques learned from audiovisual SES to improve robustness and generalizability of RadioSESNet to unseen environments and users.

We evaluate *RadioSES* using a commercial off-the-shelf (COTS) mmWave radar using synthetic and real-world data. To boost data collection for training, we build a data collection platform, and capture 5700 sentences from 19 users. Our results show that the radio modality can complement audio and bring similar improvements to that of video modality while not imposing visual privacy issues. We extensively test *RadioSES* in different number of mixtures and a variety of environmental settings. When compared to the state-of-the art audio-only method (*e.g.*, DPRNN-TasNet [84]), *RadioSES* brings around 3 dB improvements for separating noisy mixtures, along with benefits of estimating the number of sources and associating output streams. The improvements are not only in terms of SDR, but also in intelligibility and perceptual quality. Our results indicate that audioradio methods have tremendous potential for SES tasks, as they enable a low-complexity, effective, privacy-preserving alternative to audio-only or vision-based methods. *RadioSES* explores an important step in this direction and will inspire follow-up research. Some experimental results of *RadioSES* are available on our project website: `https://zahidozt.github.io/RadioSES/`

In addition to our preliminary work [52] that explores the feasibility of *audioradio* speech enhancement and separation, our main contributions in this work are:

- We propose *RadioSES*, a novel end-to-end audioradio system that jointly leverages mmWave radio and audio signals for simultaneous speech enhancement and separation.
- We introduce an audioradio deep learning framework that fuses audio signals and radio signals for multi-modal speech separation and enhancement.
- We utilize adaptive encoders for time-frequency representation, perhaps for the first time, not only for audio, but also for radio signals without relying on the commonly used spectrograms.
- We build an extensive audioradio dataset and compare *RadioSES*'s performance in various conditions with state-of-the-art methods. *RadioSES* achieves 3 to 6 dB SiSDR improvements in separating two and three-person mixtures, respectively.

The rest of the chapter follows a preliminary in Section 4.2. Section 4.3 presents an overview, with detailed design in Section 4.4 and Section 4.5. We give dataset and implementation details in Section 4.6, and present the results in Section 4.7. Last, we discuss in Section 4.8 and summarize the Chapter in Section 4.9.

4.2 Preliminary

In this section, we start with an illustration to explain what radio devices measure. Channel-impulse response (CIR) of a radio device is affected by the motion in the environment. Human vocal folds create μm level vibration displacement on the surface of the human body, especially in

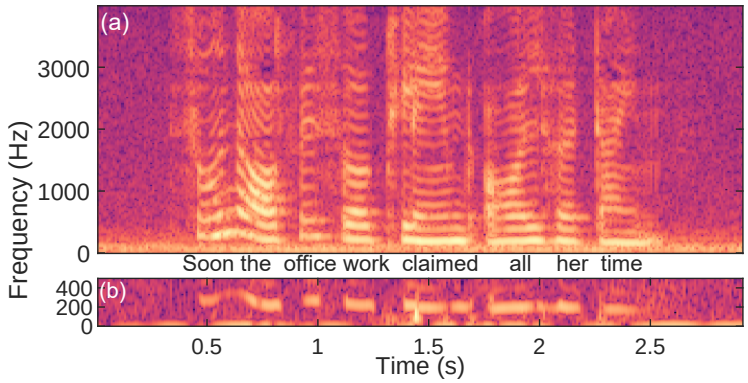


Figure 4.2: a) Spectrogram of speech, captured with a microphone and sampled at 8 kHz, b) Spectrogram of radio signal, captured from vocal fold's of the speaker in a)

the throat region, and this displacement changes the amplitude and phase of the returned complex-valued radar signals. As shown in Fig. 4.2, the low-frequency component of the radio captured spectrogram and microphone captured spectrogram are extremely similar, as the two modalities measure the same *mechanical vibration*. Radio devices potentially enable measuring voice activity (as the silence instants do not include vibration), and pitch tracking. As it will be shown later, this information from radio signals will be combined with the *corrupted* audio signals for high fidelity speech enhancement and separation. We note that, although Fig. 4.2 includes spectrograms for illustration, *RadioSES* uses learnable encoders for time-frequency representation of both audio and radio modalities.

4.3 System Overview

As an overview, *RadioSES* requires a device with mmWave sensing capabilities, and a microphone, (*e.g.* [16, 17]). The monaural microphone records ambient sound, and the mmWave radar is expected to output separate streams for each sound source, where we constrain our investigation to speech signals. Although it is possible to place radar in a separate location, we assume

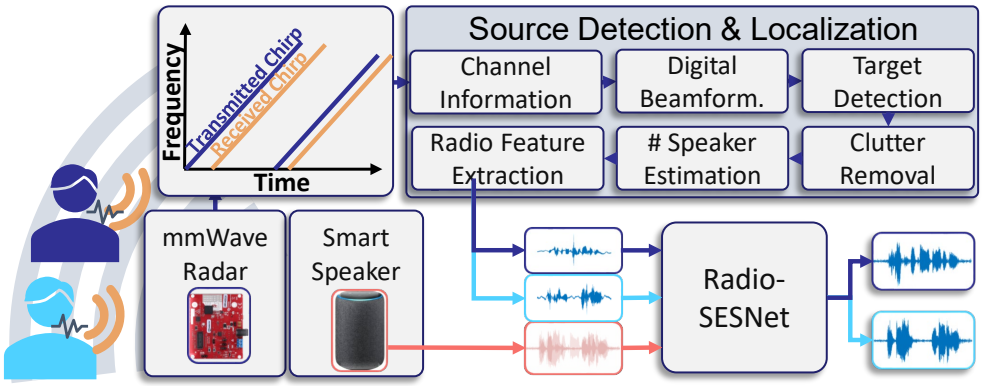


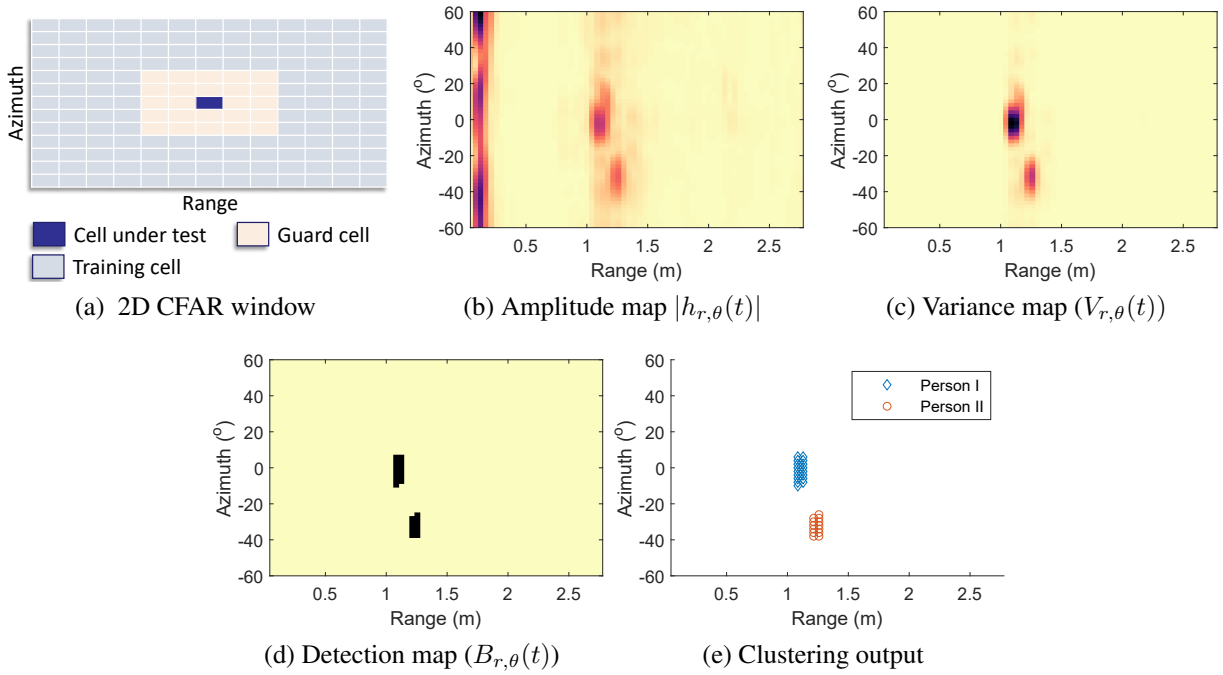
Figure 4.3: *RadioSES* Design

the radar and microphone to be colocated, as in [17]. We expect the speaking objects to be in front of the radar. In addition, although radars can sense NLOS conditions, we only investigate LOS in this work as our goal is not to eavesdrop. The application scenarios of *RadioSES* can be one or more persons speaking in front of a computer, smart hub, or a phone, with LOS.

Having speaking persons in the field-of-view (FoV), *RadioSES* detects near stationary bodies and uses the output to estimate and associate sources with the extracted sound signals. Unlike microphone arrays, using mmWave sensing enables to capture individual data streams not only from different azimuth angles, but also from varying distances. After these tasks, an efficient multimodal deep learning module is used to estimate the clean speech(es), which can be used as speech or passed through a speech-to-text engine to convert into commands[1].

The first main block of *RadioSES*, source detection and localization in Fig. 4.3, is explained briefly in Section 4.4, whereas the second block, deep learning module is further detailed in Section 4.5.

[1]Although implementing a speech-to-text conversion after a speech enhancement stage sounds plausible unless the speech enhancement stage is optimized jointly with ASR-based objectives, it can degrade performance [150].



(a) 2D CFAR window (b) Amplitude map $|h_{r,\theta}(t)|$ (c) Variance map $(V_{r,\theta}(t))$

(d) Detection map $(B_{r,\theta}(t))$ (e) Clustering output

Figure 4.4: Illustration of Sound Detection & Localization Module of *RadioSES*

4.4 Radio Feature Extraction

As shown in Fig. 4.3, the goal of the radio feature extraction module is to output individual radio streams from sources in the environment. To achieve that, we adapt a variety of methods in an efficient pipeline to detect and locate targets. Unlike existing works, such as [2,45], *RadioSES* does not rely on a spectrogram-based metric to localize people in the environment, but utilizes classical, efficient methods to extract the corresponding range-azimuth bins.

Channel Information *RadioSES* can work with any type of radar that can report a *channel impulse response* (CIR), although we use a frequency modulated continuous wave (FMCW) radar. When using an FMCW radar, extracting the CIR requires applying an operation called range-FFT, which is a common operation and we refer the reader to related work [121]. Similar to the definition in Chapter 2, we define the CIR at the m-th antenna of a multi antenna receiver as:

$h_m(\tau)$ as:

$$h_m(\tau) = \sum_{r=0}^{R-1} \alpha_{m,r}\delta(\tau - \tau_r) + \epsilon(\tau), \tag{4.1}$$

where R is the number of the CIR range bins, $\delta(\cdot)$ is the Delta function representing the presence of an object at the corresponding location, $\alpha_{m,r}$ and τ_r denote the complex amplitude and the propagation delay of the r-th range bin, and ϵ denotes the additive noise, respectively. Here, the range resolution ΔR can be inferred from the time resolution, $\Delta\tau$, which is inversely proportional to bandwidth (corresponding to $4.26cm$ for our device). Therefore, a separate stream from very close targets can be extracted. The CIR in (4.1) is captured repeatedly during sensing and is time dependent. To simplify (4.1), we denote the CIR from m-th antenna, at r-th range bin, at time index t as $h_{m,r}(t)$. Note that, $h_{m,r}(t)$ is quantized with respect to time, range bin, and antenna index.

Digital Beamforming Using the individual received streams from each antenna, *RadioSES* extracts range-azimuth information with classical beamforming [151]. Range-azimuth CIR is denoted by $h_{r,\theta}(t)$, where θ represents the azimuth angle. Since our virtual antenna array elements are placed $d = \lambda/2$ apart, where λ is the wavelength, $h_{r,\theta}(t)$ can be given as:

$$h_{r,\theta}(t) = \mathbf{s}^H(\theta)\mathbf{h_{m,r}}(\mathbf{t}) + \epsilon(t), \tag{4.2}$$

where $\mathbf{s}^H(\theta)$ is the steering vector for angle θ, and ϵ is the additive noise. The coefficients of the steering vector are:

$$s_m(\theta) = \exp\left(-j2\pi\frac{d\sin\theta}{\lambda}\right), \tag{4.3}$$

and the channel vector is $\mathbf{h_{m,r}(t)} = [h_{1,r}(t), h_{2,r}(t), ..., h_{M,r}(t)]$, with M being the total number of antenna elements.

Target Detection To detect human bodies in the environment, *RadioSES* first extracts the reflecting objects in the environment. As suggested by (4.1), the presence of objects creates strong returned signals, whereas when there is no object, returned signals only consist of noise. For target detection, we utilize a classical approach in the radar literature, constant false alarm rate (CFAR) detector [152], which adaptively estimates the background noise for different bins and thresholds each range-azimuth bin accordingly. As shown in Fig. 4.4a, the 2D CFAR window is denoted with C, and CFAR threshold is denoted with γ. This window is applied to the magnitude of the range-azimuth plane, and the corresponding range-azimuth plane is shown in Fig. 4.4b. Therefore, the CFAR detection rule on the range-azimuth plane is given as:

$$B_{r,\theta}^{\text{CFAR}}(t) = \mathbb{1}\{(C \star |h_{r,\theta}|)(t) > \gamma(|h_{r,\theta}(t)|)\}, \tag{4.4}$$

where $\star$ and $\mathbb{1}\{\cdot\}$ denote the convolution operation and indicator function, respectively.

Clutter Removal The previous module extracts a binary map with bins with reflecting objects, which can include static objects. On the other hand, even when a person is stationary, the radar signal still captures a variation at the person's location, due to inherent body motion from breathing and heart rate, a phenomenon used extensively in mmWave based person detection [20, 153]. Therefore, to remove the static objects and detect human bodies, we extract the variance at each range-azimuth bin, and use a threshold to identify static objects. We denote the variance of $h_{r,\theta}(t)$ with $V_{r,\theta}(t)$, where an example can be seen in Fig. 4.4c. Therefore, human detector output is $B_{r,\theta}^{\text{stat}} \triangleq \mathbb{1}\{V_{r,\theta}(t) > \text{H}^{\text{stat}}(r,\theta)\}$. Furthermore, bodies

with excessive motion can also be filtered using a similar approach, and we reject those by: $B^{\text{mov}}_{r,\theta} \triangleq \mathbb{1}\{V_{r,\theta}(t) < \text{H}^{\text{mov}}(r,\theta)\}$, where $\text{H}^{\text{stat}}(r,\theta) \triangleq \frac{\eta^{\text{stat}}\cos(\theta)}{(1+r\Delta R)^2}$, $\text{H}^{\text{mov}}(r,\theta) \triangleq \frac{\eta^{\text{mov}}\cos(\theta)}{(1+r\Delta R)^2}$, η^{stat} and η^{mov} are empricially found thresholds. The minimum and maximum variances are defined with respect to (r,θ), in order to accommodate changing reflection energy with respect to angle and distance. The resulting binary detection map, $B_{r,\theta}(t)$ is found by extracting intersection of all binary maps, i.e. $B_{r,\theta}(t) = \{B^{\text{CFAR}}_{r,\theta} \cap B^{\text{stat}}_{r,\theta} \cap B^{\text{mov}}_{r,\theta}\}(t)$, as shown in Fig. 4.4d.

Number of People Estimation Each bin of binary detection map, $B_{r,\theta}(t)$ spans $(\Delta R, \Delta\theta)$ distance in 2D space. Considering the high range and angular resolution, a human body can span multiple bins in $B(r,\theta)$. To estimate the number of people, *RadioSES* clusters binary detection maps using a non-parametric clustering method, DBSCAN [154]. The parameters for DBSCAN are set empirically, and an example clustering is shown in Fig. 4.4e. Furthermore, since the number of people estimation and center extraction is done repeatedly for a window of size W, there is a need to match the locations of bodies at different time indices. We use Munkres' algorithm [155] to continuously track the location of users.

Radio Feature Extraction Having extracted the number of persons and the corresponding range-azimuth bins, *RadioSES* extracts the complex radar signals from each person's center directly, following recent raw-data-based approaches [156]. As there are many range-azimuth bins associated with the same person, *RadioSES* extracts the median bin for testing, whereas multiple nearby bins are used for training, which helps to boost dataset size and mitigate overfitting. Output dimensionality of the radar signals is 2×1000 at 16bits for a 1-second stream, which is lower than the microphone and typical video streams.

4.5 Audioradio Deep Learning Model

In this section, we explain the structure of the deep learning model used in *RadioSES*, named RadioSESNet. We first introduce the relevant background in SES and our design rationale in Section 4.5.1 and then detail our design in Section 4.5.2.

4.5.1 Background and Design Rationale

Background: Usually, an SES model follows the architecture in Fig. 4.5, with an encoder, masker, and a decoder block [157]. Input encoding is multiplied with an estimated mask, which uses a decoder to reconstruct the time-domain signal. Early works have used STFT as the encoder, with the ideal binary mask being the training objective [158]. The performance can be increased by using more optimal masks (such as complex ratio mask [79]); however, these still suffer from the fact that STFT-based encoding is not necessarily optimal for speech separation, and methods that replace STFT with adaptive encoders are found to be more optimal [83].

Design Rationale: *RadioSES* uses the same structure as in Fig. 4.5, with the addition of a radio stream. Radio streams are encoded, and concatenated with the audio stream to estimate the masks. However, this involves a few design choices as follows: Unlike audio signals, radio signals are complex-valued and both real and imaginary parts change with respect to the motion and vibration [3, 156]. If a spectrogram representation is used as an input, not only it may not be optimal for the training task, but it usually involves throwing away some signal content by only extracting amplitude, or half of the spectrogram (*e.g.* only positive Doppler shifts), as in [48, 51]. Using either the real or imaginary part of the signal (as in [2]) or combining both parts optimally with a linear projection [45] also loses important signal content. Based on this, *RadioSES* uses

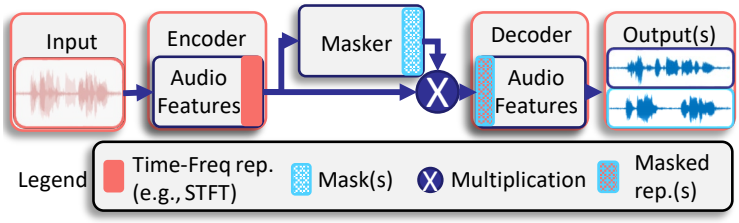


Figure 4.5: Typical SES System Workflow

adaptive front-end for radio streams.

To make *RadioSES* work with raw radio inputs, we apply random rotation in IQ plane, as proposed by previous work [156][2]. However, unlike [156], we apply a high-pass filter on returned signals to reduce the effect of body motion. The high pass filter is needed for RadioSESNet to run with raw radar inputs, as will be shown in Section 4.7. We select the cutoff frequency of the high pass filter at 90 Hz in order not to filter vocal folds harmonics. Afterward, the radio signals are encoded with an adaptive encoder, as explained in Section 4.5.2.

After the encoder, we process audio and radio streams separately with individual blocks to exploit long-term dependencies within each modality. To that end, we process each modality through an efficient dual-path RNN block (DPRNN). DPRNN blocks do not suffer from limited context, a well-known issue with fully convolutional models [84]. Afterward, we combine two modalities via resizing and concatenation on the feature dimension. These models are further processed with DPRNN blocks and 1D decoders before outputs.

4.5.2 RadioSESNet Design

[2]We refer the reader to [3] for IQ representation of the returned signals, and to [156] for discussion and introducing random rotation.

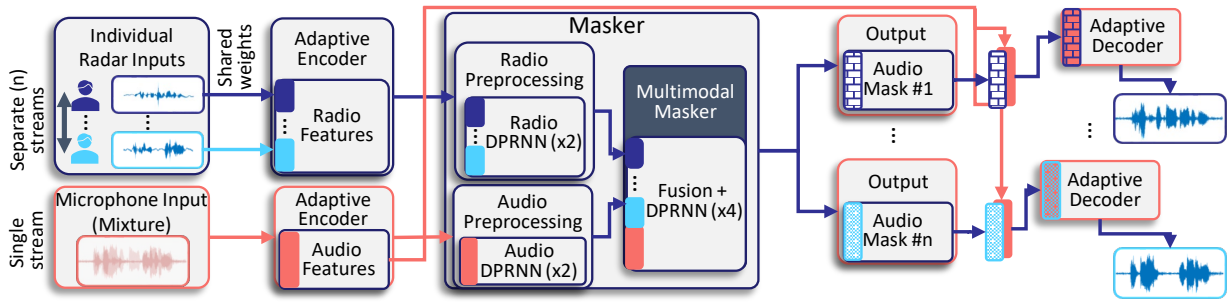


Figure 4.6: RadioSESNet Structure

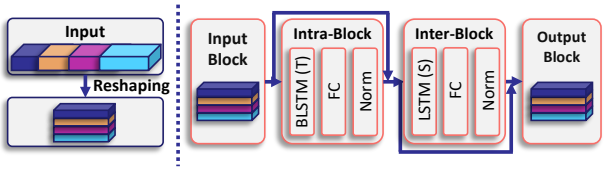


Figure 4.7: Left: Reshaping operation with overlapping windows. Right: Single DPRNN Block

4.5.2.1 Encoders

The audio encoder of RadioSESNet consists of a 1D convolutional layer, with kernel size 16, and number of kernels 256, followed by ReLU nonlinearity and layer normalization. Radio channel uses another 1D convolutional layer, nonlinearity, and normalization. These layers have the same parameters, except the number of filters being 64, due to the lower sampling rate. Stride size is set to 1/2 of the kernel width, resulting in 50% overlap between convolutional blocks. After the first layer, a second 1D convolution reduces the dimensionality to 64 for audio, and 16 for radio. Each radio stream uses the same encoder block to create an STFT-like representation. We denote the distorted input audio with $\tilde{a}$, and radio streams with r_i, where i denotes the i^{th} radio stream. Output of the audio and radio encoders are represented with $\mathbf{X}_\star \in \mathcal{R}^{N_\star \times L_\star}$, with $\star \in (a, r)$, for audio and radio stream, where we drop the index i for simplicity. Here, $N_\star$ represents the number of features, and $L_\star$ represents the number of time samples of encoded representation.

4.5.2.2 Masker

Both encoded modalities are combined to estimate the masks for each source, as illustrated in the masker of Fig. 4.6. Each modality passes through individual DPRNN blocks, then fused by vector concatenation, and passes through four more DPRNN blocks before estimating the mask with a 2D convolutional layer, which matches the output with the expected mask number and size.

DPRNN Processing: For processing the encoded data, we use DPRNN blocks [84], where an example DPRNN workflow is presented in Fig. 4.7. DPRNN processing consists of reshaping the input data to a 3D representation, through means of extracting overlapping blocks, and concatenating through another dimension, and applying two consecutive RNN layers to different dimensions of the input block. The output of the reshaping operation can be represented as $\hat{\mathbf{X}}_{\mathrm{a}} \in \mathcal{R}^{N_{\mathrm{a}} \times K_{\mathrm{a}} \times S_{\mathrm{a}}}$, with K_{a} and S_{a} denoting the block length and number of blocks. The input, output representations $\mathbf{X}_{\mathrm{r}}$, $\hat{\mathbf{X}}_{\mathrm{r}}$ and dimensionalities N_{r}, L_{r}, K_{r} and S_{r} are defined similarly for radio channel, and given in Table 4.1, whereas the flow for a single DPRNN processing is given in Fig. 4.7.

After a suitable reshaping operation, the input blocks are fed to an RNN module, which is operated along the S dimension of the 3D input, followed by a fully connected layer, and layer normalization. After a skip connection in between, a similar operation is repeated through K dimension to capture larger distance relationships between blocks. Each RNN block has depth

Table 4.1: Parameters for the Masker Layer for 2-Mix

Audio	N_{a} 64	K_{a} 128	S_{a} 48
Radio	N_{r} 16	K_{r} 16	S_{r} 48
Concatenation	N_{c} 96	K_{c} 128	K_{c} 48

1, and fully connected layers are used to match the input size to the output size, which enables concatenating DPRNN blocks without any size mismatches.

4.5.2.3 Decoder

At the output of the masker, the number of masks equal to the number of people are estimated, which is then used to *decode* the signal to extract time domain audio signals. DPRNN blocks are converted back to a representation similar to the one at the input, by an overlap-add method [84]. The signal is fed through the decoder, which applies a transposed convolution operation. The output is a single channel representation, with the same dimensionality and the same number of filters in the encoder to preserve symmetry, and it is also adaptive.

4.5.2.4 Training

In order to train RadioSESNet, we use scale-invariant signal-to-distortion (SiSDR [159]) as the loss function between the time-domain signals, which is given by:

$$\mathbf{SiSDR}(\mathbf{a}, \hat{\mathbf{a}}) = 10 \log_{10} \left(\frac{|| \frac{\hat{a}^T a}{||a||^2} a ||}{|| \frac{\hat{a}^T a}{||a||^2} a - \hat{a} ||} \right), \tag{4.5}$$

where a and $\hat{a}$ denote the target and the estimated sound signals. Use of SiSDR prevents scaling effects to dominate the error calculation, as the amplitude of extracted speech is not of interest . The SiSDR loss has been combined with L_2 norm regularization on the weights, where the decay factor is set to $1e^{-6}$. Since a separate model for different numbers of users has been trained, *RadioSES* switches to the appropriate model by estimating the number of sources.

4.5.2.5 Other Design Considerations:

Complexity and causality are particularly considered in our design.

Complexity: RadioSESNet has a compact design, with only 2.1M parameters. Among these, the radio stream occupies 320k parameters, which could easily be fit on a small device. Forward pass of a 3-second input with RadioSESNet takes 4ms on a modern GPU with batch processing, which is only 0.4ms slower than the corresponding audio-only method.

Causality: RadioSESNet uses unidirectional LSTMs in the recurrent layers of inter-block processing, whereas intra-blocks rely on BLSTMs which requires having the complete block in S dimension. Therefore, RadioSESNet can work causally, with roughly $150ms$ delay. We leave the investigation of a real-time work to future, but *RadioSES* is already close to real-time processing, unlike [63, 93].

4.6 Experiment and Implementation

4.6.1 Data Collection

Hardware: We build a data collection platform, as seen in Fig. 4.8, to obtain large-scale data to train, validate, and evaluate *RadioSES*. As extracting clean and non-reverberant ground truth samples are important, we reduce the echo in the room by sound-absorbing pads. We collect clean audio data with a Blue Snowball iCE microphone, sampled at 48 kHz, radar data using a Texas Instruments (TI) IWR1443 mmWave radar, and video data using the front-facing camera of an iPhone 11 Pro. The radar is set to operate with a bandwidth of 3.52 GHz at a sampling rate of 1000 Hz. We align the radio signal and audio signal in the time domain using the correlation

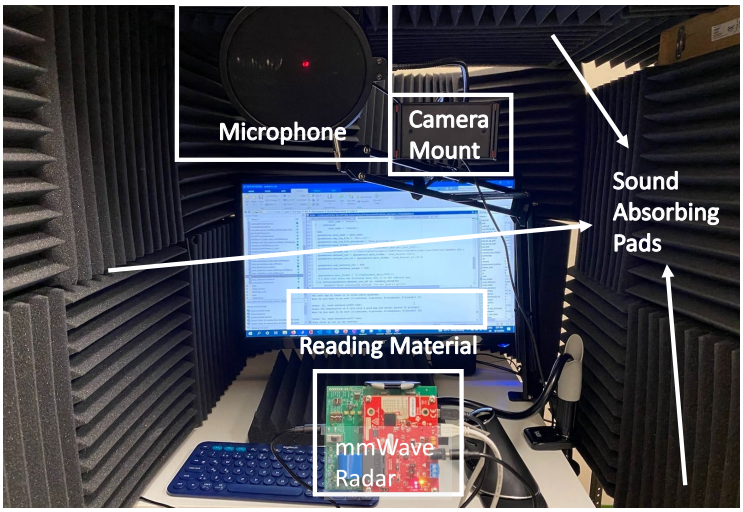


Figure 4.8: Setup of Data Collection Center

of their energy. Video data, captured at 1080p and 30 fps, is collected for future research and not used in this work; although the accompanying audio files are used for training.

Speaker setting: We recruit 19 users including native speakers and speakers with different accents to read phonetically rich sentences from the TIMIT corpus [160]. Our speakers come from a diverse background, where there are 5 native English speakers, along with 9 Chinese, 2 Indian, 2 Turkish, 1 Korean accented speakers. We remove sentences that are shorter than 25 characters in the dataset. Since the size of TIMIT corpus is limited, 200 common and 100 unique sentences are read by each participant. A total of 5700 sentences were read, including 2100 **unique** sentences and 5762 **unique** words. The sentences are presented in mixed order, and our dataset includes a lot of pauses and filler words, in contrast to publicly available datasets, which usually include professional speakers (*e.g.* LibriMix [161]). During data collection, users sit approximately 40cm away from the radio device and read each material at a normal speaking volume while not moving excessively.

Data generation: To generate the noisy and mixture sound signals, we follow the recipe used in LibriMix [161] with the noise files from WHAM dataset [145]. We randomly select 13 users for training, and 4 users (2 male, 2 female) for evaluation. The validation set includes the remaining two users, and unused speech of the users in the training set. After downsampling all audio files to 8kHz, we create synthetic mixtures based on the shortest of the combined files, with a minimum duration constraint of 3-seconds. Each user's recordings are repeated ten times on average, which results in 25,826 utterances (≈30 hours). The gain factors are found by normalizing the loudness of speech and noise signals, and creating noisy mixtures in $[-5, 5]$ dB signal-to-noise rate (as in [161]). We create two evaluation sets:

- **Seen:** mixtures from seen users, but unheard sentences (*a.k.a* closed-condition)
- **Unseen:** mixtures from unseen users (*a.k.a* open-condition)

This helps us to better understand the dependency on seen/unseen users in *RadioSES*, as different users' radio signals can be different, not only due to their speaking, but also due to their body motion and physical characteristics. Other experimental settings are also introduced and investigated in Section 4.7. On the other hand, our experiments include mostly overlapping speech, to better illustrate the difference between audio-only and audioradio methods, and we leave evaluation of partially overlapping speech to future for conciseness.

Dataset Considerations for Improving Robustness: A multimodal system can fail easily and focus to use a single modality, which is known as mode failure. To prevent this and to further improve robustness, our dataset creation procedure includes the following:

- **Same-speaker mixtures:** Our dataset includes same-speaker mixtures, in order to prevent mode failure, which is shown to be effective in the audiovisual domain [162].

- **Multi-microphone mixtures:** As our data collection procedure includes two microphones, we randomly select one when generating each mixture. Our evaluation is done with the better microphone (Blue), but this also boosts dataset size multiple folds without collecting more data.

- **Clean and Noisy Mixtures:** Unlike the LibriMix dataset [161], we create both noisy and clean mixtures of multiple speakers and use them to train a single model. Therefore, *RadioSES* uses a single model, whether an environment is clean or noisy.

4.6.2 Implementation Details

We implement data collection and raw data processing modules of *RadioSES* in MATLAB, whereas the deep learning model is implemented in PyTorch, with the help of Asteroid library [163] to follow standard training and evaluation protocols in monoaural SES, and to borrow implementations of existing methods, such as ConvTasNet [83] or DPRNNTasNet [84]. We train RadioSESNet and DPRNNTasNet for 60 epochs, using a starting learning rate of $1e^{-3}$, which is halved when the validation loss did not improve for 5 consecutive epochs. Furthermore, the learning rate is scaled by 0.98 every two epochs, as in [84]. An early stopping criterion is set to 15 epochs. To accelerate training, we use mixed-precision training. Thanks to the low complexity design of RadioSESNet, a single epoch takes roughly 10 minutes to train, with a batch size of 24, using a single NVIDIA RTX 2080S GPU.

Considerations to Improve Robustness: As noted previously, although microphone signals mostly correspond to speech signals, radar signals can be affected by motion, vibration, and environmental factors. Furthermore, it is usually not straightforward to make a multimodal system

work easily. To improve the robustness of radio signals, we implement the following:

- **Capturing Multiple Snapshots** Since a single user spans multiple range-azimuth bins due to high resolution, we record multiple range-azimuth data in our dataset. In each epoch, we randomly select a range-azimuth bin for training among a maximum of 8 candidates, whereas validation and testing use the median bin. This boosts the dataset size significantly without relying on synthetic methods and enables to use a wider range of bins, instead of searching for the most optimal bin.

- **Input Distortions:** The input radio streams are distorted in different ways. These include introducing random rotation [156], adding noise at different variance levels, replacing some part of the radio signal with zeros (to imitate data loss), or removing some radio signals completely, as suggested by [164] to reduce mode failure.

4.7 Evalution

In this section, we introduce performance metrics and baselines for comparison, which are followed by results using *RadioSES*. Afterward, we investigate the practical limits and robustness of *RadioSES* by analyzing environmental effects. Next, we present a real-world case study to illustrate the benefits coming from *RadioSES*. Last, we evaluate *RadioSES* in some interesting cases, such as noisy, partial inputs, and conduct an ablation study.

Performance Metrics: We report the following metrics to evaluate the performance of *RadioSES*:

- SiSDR [159]: Scale-invariant signal-to-noise ratio, which is an indicator of signal levels, with a normalization factor to prevent scaling of the signals to increase metric unfairly.

- SIR: [165]: Signal-to-interference ratio, which measures the leakage from one person to another when there are multiple speakers, and only reported for SS tasks.

- STOI [126]: Short time intelligibility metric, correlates with the word error rate, reported from 0 to 1.

- PESQ [125]: Perceptual evaluation of the sound quality, measured from 0 to 5. Since measuring human perception requires user studies, this metric has been proposed as an alternative, when user studies are not feasible.

Baseline Methods: We include several radio-only and audio-only methods in the literature for a variety of tasks. First, as a radio-only method, we implement WaveVoiceNet in WaveEar [48]. This approach uses the radio modality alone to (re)construct sound signals from vocal folds vibration, and assumes no available microphones. It reconstructs the magnitude of audio spectrograms and uses Griffin-Lim based phase reconstruction. We use the oracle phase of the clean audio signal instead, which poses an upper limit on its performance. Another recent work [51] is similar to our work in combining two modalities, yet their end-to-end system focuses on translating single speaker noisy voice commands into text without a sound output and not comparable to our method.

We compare the performance of *RadioSES* with other audio-only baselines, to illustrate gains from radio modality, and sustained performance of *RadioSES*. We include ConvTasNet [83], one of the first adaptive-encoder based systems that outperformed STFT-based masks. Second, we include DPRNNTasNet, which is the audio-only baseline of *RadioSES*. DPRNNTasNet has shown to outperform ConvTasNet significantly and can be considered as one of the state-of-the art methods. Third, we use SudoRMRF [166], which simplifies DPRNNTasNet by replacing

Table 4.2: Results for enhancing single speaker speech. Seen: closed-condition, and unseen: open-condition

Evaluation	Seen			Unseen		
Model	SI-SDR	STOI	PESQ	SI-SDR	STOI	PESQ
Input	3.9	0.74	1.55	3.8	0.70	1.54
WaveVoiceNet	0.6	0.60	1.28	0.7	0.62	1.27
ConvTasNet	**14.5**	0.90	2.67	**13.6**	**0.87**	**2.55**
SudoRMRF	14.0	0.88	2.32	12.2	0.84	2.04
DPRNNTasNet	14.2	0.89	2.62	13.0	0.86	2.46
RadioSES	14.5	**0.90**	**2.68**	13.3	0.87	2.52

the RNN blocks with downsampling and upsampling blocks and is shown to achieve similar performance.

Last, we cannot compare with UltraSE [63], as it uses ultrasound modality, and different speakers and noise dataset. Due to changes in datasets and different sampling rate (16 kHz), it is not possible to copy their results and draw a direct comparison. On the other hand, UltraSE performs similar to ConvTasNet in 2-person mixtures, which we have included as a benchmark in our study.

4.7.1 Speech Enhancement

In speech enhancement, *RadioSES* brings improvements to the audio-only baseline methods, as shown in Table 4.2. Since the background signals are statistically different than speech signals, we see relatively small improvements. This observation is consistent with audiovisual methods (*e.g.* 0.1 dB improvement in [93]), and shows that *RadioSES* learns to exploit the radio information. On the other hand, results from WaveVoiceNet suggest that the radio modality is not sufficient to (re)construct less-noisy audio, and may not be feasible within our experimental setting. This can be attributed to differences in the hardware (special hardware is used in [48]),

our phonetically rich diverse dataset (5762 unique words vs. 631 in [48]), and users. As the results are poor, we do not investigate WaveVoiceNet further in our experiments. Performance of *RadioSES* matches to that of ConvTasNet, with certain qualitative differences, such as 1.5s look-ahead in ConvTasNet, and higher computational complexity. We also note that, our implementation uses a pretrained ConvTasNet on a much larger dataset, which potentially improves the overall performance. This section investigates the case, where the background is non-speech noise. Having an interfering speech signal can also be considered as part of the speech enhancement problem, yet the enhancement methods usually require some prior information to focus on the particular speech. If such prior information does not exist, it is more reasonable to evaluate the performance against speech separation models. In order to have a fair comparison, we evaluate this case in the following sections, under speech separation.

4.7.2 Speech Separation

In this section, we present the speech-separation results with *RadioSES*, along with the previously mentioned baselines in Table 4.3. For both separating single and noisy speech tasks, *RadioSES* outperforms a variety of state-of-the-art methods in audio-only domain, including DPRNNTasNet. Our DPRNNTasNet implementation achieves 13.5 SiSDR in 2-person clean mixtures, which is close to the reported value in the LibriMix dataset, 16.0. Significant improvements with respect to SIR can be observed in both clean and noisy cases, which indicates the usefulness of the radio channel for separating the mixtures, and suppressing the interference. Furthermore, even though there is more variety in radio inputs (*e.g.* radio channel inputs are not only affected by the sound, but also by ambient motion and physical characteristics), *RadioSES* can still gener-

Table 4.3: Evaluation in 2-Person Mixtures (SS)

		2-person mix (clean)				2-person mix (noisy)			
	Model	SI-SDR	SIR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ
Seen	Input	0.2	-0.4	0.71	1.71	-1.7	0.3	0.61	1.37
	ConvTasNet	11.3	18.5	0.87	2.53	6.1	16.8	0.77	1.78
	SudoRMRF	10.9	15.4	0.84	2.60	4.7	16.4	0.68	1.77
	DPRNN	13.5	21.5	0.91	2.63	8.9	20.3	0.81	1.96
	RadioSES	**15.4**	**23.6**	**0.94**	**2.83**	**10.9**	**23.3**	**0.85**	**2.10**
Unseen	Input	0.0	0.53	0.70	1.62	-1.8	0.30	0.60	1.39
	ConvTasNet	9.5	16.0	0.84	2.38	5.2	15.0	0.72	1.67
	SudoRMRF	6.2	11.5	0.76	2.13	1.0	13.0	0.60	1.39
	DPRNN	10.8	18.1	0.86	2.38	7.0	17.3	0.75	1.83
	RadioSES	**14.5**	**22.3**	**0.92**	**2.70**	**10.3**	**22.5**	**0.83**	**2.05**

Table 4.4: Evaluation in 3-Person Mixtures (SS)

		3-person mix (clean)				3-person mix (noisy)			
	Model	SI-SDR	SIR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ
Seen	Input	-3.2	-2.8	0.60	1.37	-4.2	-2.8	0.55	1.30
	DPRNN	7.2	14.0	0.81	1.95	4.9	15.7	0.74	1.68
	RadioSES	**11.6**	**19.4**	**0.88**	**2.31**	**9.3**	**19.2**	**0.83**	**1.96**
Unseen	Input	-3.2	-2.8	0.58	1.37	-4.2	-2.8	0.54	1.31
	DPRNN	4.2	10.2	0.73	1.72	2.6	12.5	0.66	1.55
	RadioSES	**10.7**	**18.2**	**0.86**	**2.21**	**8.6**	**18.2**	**0.81**	**1.90**

alize better to unseen users, where the basic DPRNNTasNet suffers. *RadioSES* not only improves signal metrics, but also intelligibility and the perceptual quality metrics (PESQ). The difference between the audio-only baseline becomes larger, especially when the input mixtures are corrupted with noise and when there are multiple people. We also train *RadioSES* with three people mixtures. As shown in Table 4.4, the improvements from *RadioSES* is even greater for 3-person mixtures, as the radio stream helps to extract individual streams from each user. Since the performance gains from *RadioSES* increase with more users, we expect it to work well for 4 or more users. We do not test those cases for brevity.

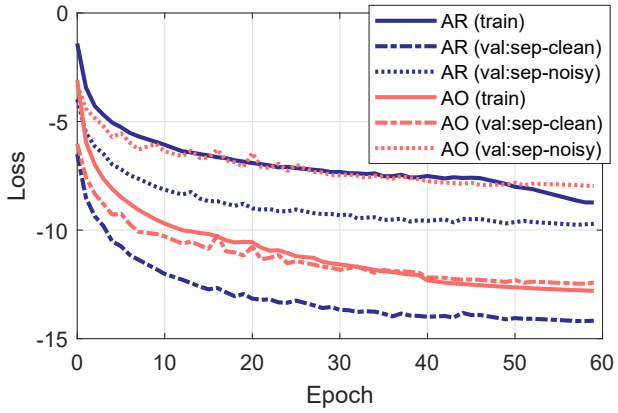
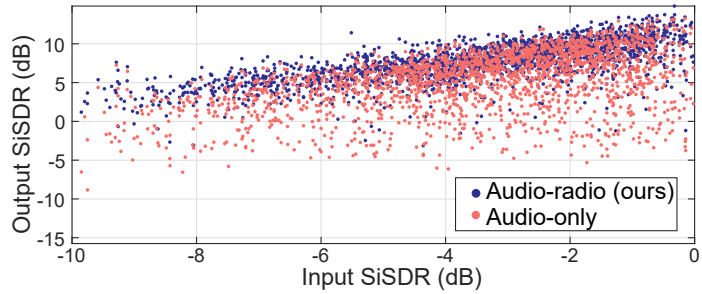


Figure 4.9: Learning curve for audio-only (AO) and audioradio (AR) for separating 2-person mixtures

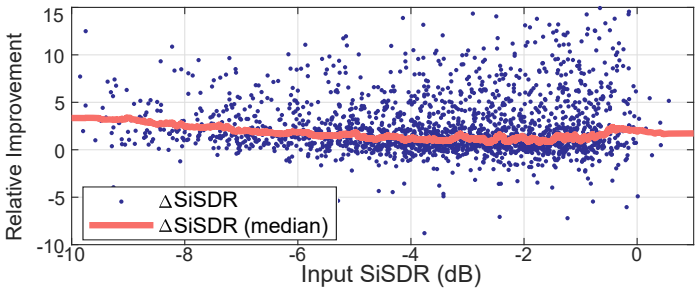
4.7.3 Comparison with Audio Only Baselines

As mentioned previously, introducing another modality has many benefits, such as guiding the loss function at the beginning of training to solve the permutation problem and estimating the number of sources. To that end, in Fig. 4.9, we compare the loss values on training and validation sets. As shown, the audioradio system has a much steeper learning curve at the beginning, along with a better convergence point.

Furthermore, in Fig. 4.10a, we compare the output SiSDR of *RadioSES* with its audio-only baseline. As shown, our proposed method is superior to the audio-only baseline, and the performance gains are consistent through different input SiSDR levels. To investigate the consistency of audioradio system over audio, we plot the differential gain in terms of SiSDR in Fig. 4.10b from the radio channel. To characterize the incorrect associations, we check the amount of samples with $\Delta(DB_i) < -3$ is 1.03%, indicating correct physical association of sources for 98.97% of the time.

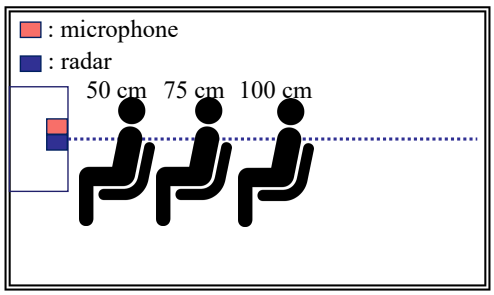


(a) Output of sound separation

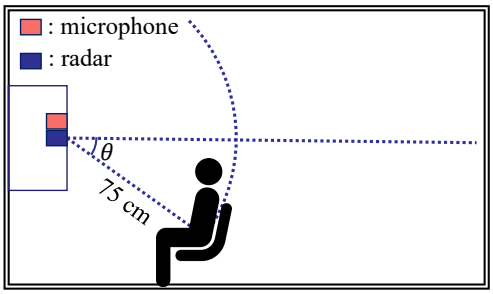


(b) Relative gains from radio channel

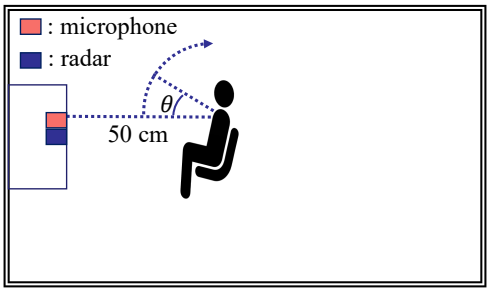
Figure 4.10: Comparison of *RadioSES* with audio-only baseline in 2-person noisy mixture



(a) Distance setting



(b) Incident angle setting



(c) Head orientation setting

(d) Multi-user setting

Figure 4.11: Multiple experimental settings

4.7.4 Impact of Experiment Setting

We further evaluate the performance of *RadioSES* in varying settings, conducted in a different location than the original data collection location. Since it is difficult to *simulate* the extracted radio signals from different environmental scenarios, we collect data at a variety of settings. For example, to test the effect of distance, we collect multiple user data at different distances, (e.g. 75cm), and create mixtures from that location. We normalize input data streams to the same loudness levels for a fair comparison, although minor differences between each setting is inevitable. In order to show improvements, we present each settings' performance along with the audio-only baseline, and show how *RadioSES* preserves a better performance in those settings. For presentation, we refer *RadioSES* as the audioradio (AR) method, whereas baseline DPRNNTasNet is noted as audio-only (AO) method. As shown, *RadioSES* mostly outperforms audio-only baseline with 4dB improvement in our dataset, which includes unseen and same-speaker mixtures. This evaluation is done with clean mixtures for consistency, although we have observed similar gains in noisy mixtures as well.

4.7.4.1 Distance

First, we evaluate the effect of distance on the signal separation tasks, as illustrated in Fig. 4.11a. As shown in Table 4.5, *RadioSES* can work robustly until the speakers are 1m away from the device, and preserve the gains compared to the audio-only baseline. The performance for both cases decrease, which is due to training dataset being captured from a short distance only. As the distance increases, the received audio signals change due to the room impulse response and microphone nonlinearity, which is a phenomenon used for coarse source distance estimation with

Table 4.5: Performance with respect to distance

Exp	Distance					
Case	50 cm		75 cm		100 cm	
Metric	AO	AR	AO	AR	AO	AR
SiSDR	6.3	10.9	3.8	8.6	2.3	4.3
SIR	12.5	18.3	9.9	15.6	8.7	9.8
STOI	0.83	0.93	0.79	0.90	0.74	0.81
PESQ	2.17	2.61	1.97	2.42	1.79	2.00

microphones recently (*e.g.* [167, 168]). We note that the performance gains from radio channel do not decrease much from 0.5m and 1m, and the main bottleneck for lower performance is the variety of audio data. A high-performance system can be built by capturing more diverse **audio** data.

4.7.4.2 Orientation

Second, we ask the users to sit 0.75m away from the device and change their orientation to explore the practical area of sensing, as illustrated in Fig. 4.11b. We realize that *RadioSES* can work until $45°$, without any performance decrease, as presented in Table 4.6. The gains from the audioradio system are consistent (*e.g.* $\sim$4dB in SiSDR) through each setting, showing the effectiveness in modeling of the radio stream. Furthermore, this observation is consistent with that of distance, as a different deviation angle from the microphone does not create any distance-based nonlinearity, although it reduces the radio-reflection SNR.

4.7.4.3 Head Orientation

Third, we ask users to sit at 0.5m, and rotate their heads from 0 degrees to 15 and 30 degrees, as shown in Fig. 4.11c. For example, if a user sits in front of a laptop or monitor, they

Table 4.6: Performance with respect to orientation

Exp	Orientation							
Case	0°		15°		30°		45°	
Metric	AO	AR	AO	AR	AO	AR	AO	AR
SiSDR	3.8	8.6	3.6	7.8	4.4	8.3	4.2	8.2
SIR	9.9	15.6	9.6	14.8	10.6	15.1	10.2	15.6
STOI	0.79	0.90	0.78	0.89	0.79	0.89	0.78	0.88
PESQ	1.97	2.42	1.91	2.32	2.00	2.33	2.02	2.37

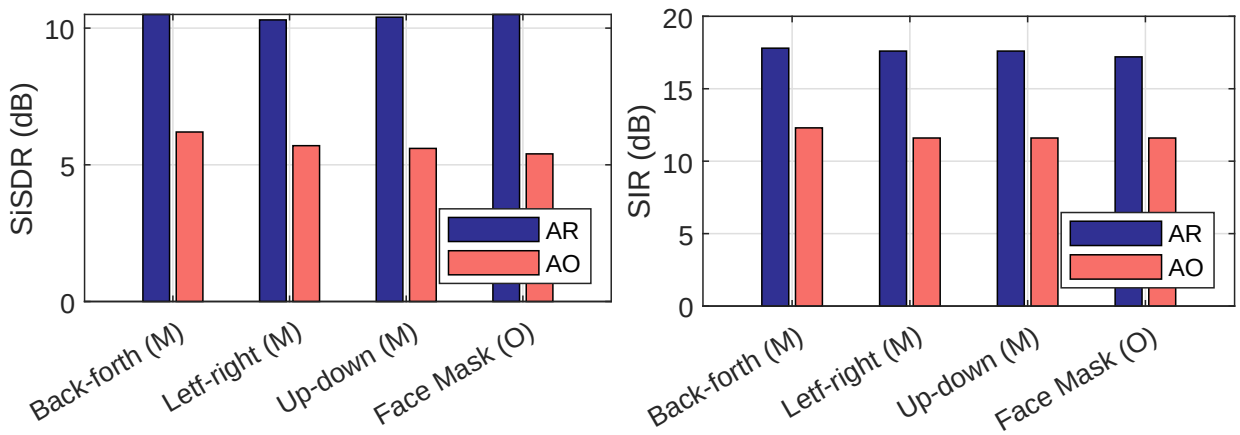
Table 4.7: Performance with respect to head orientation

Exp	Head Orientation					
Case	0°		15°		30°	
Metric	AO	AR	AO	AR	AO	AR
SiSDR	6.3	10.9	5.6	9.8	5.4	9.3
SIR	12.5	18.3	11.7	16.8	11.5	16.3
STOI	0.83	0.93	0.80	0.90	0.79	0.89
PESQ	2.16	2.61	2.11	2.46	2.10	2.43

would naturally swing their head to see different content on the screen and 30 degrees of head rotation at 0.5m enables them to see the entire area of a big screen. Furthermore, if *RadioSES* is using lip motion, instead of vocal folds vibration, we would expect the results to deteriorate quickly. The results are presented in the head orientation column of Table 4.7, which indicates that *RadioSES* is robust to changes in head orientation, even though the training procedure does not include explicit head-rotation data.

4.7.4.4 Distortion

Fourth, we ask users to perform a variety of distortions. First, we ask users to perform motions in front of the radar while speaking. To have the experiments controlled, we ask the users to move their heads up and down, left-to-right and back-and-forth naturally, as it can happen during speech. Next, we collect data with users wearing a mask, which plays a role as an occlusion.



(a) SiSDR (b) SIR

Figure 4.12: Performance when there is motion (M) of the user, or occlusion (O).

Table 4.8: *In the Wild* Experiment Results

Case	Speech Enhancement			Speech Separation		
Metric	Clean	AR	Noisy	Clean	AO	AR
WER	14	45	63	20	61	55
CER	8	32	54	11	50	40

As shown in Fig. 4.12, *RadioSES* is not affected by the head motion. Furthermore, unlike certain visual enhancement methods which lose their advantage with occlusions (as noted in [169]), *RadioSES* is robust against wearing a mask and can preserve the improvements compared to the audio-only method. This is due to the fact that vocal folds vibration are extracted from the body and throat, not from the face. Similar improvements with respect to STOI (*e.g.*from 0.8 to 0.9), and PESQ (*e.g.*from 2.1 to 2.5) are also observed, but not reported in the figures.

4.7.5 Case Study in the Wild

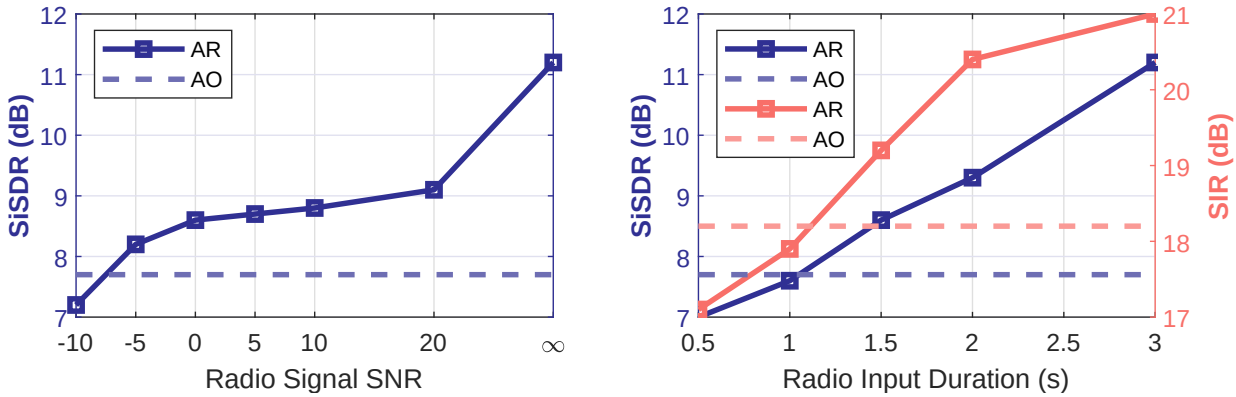
In this experiment, we ask multiple users to sit within the same room, and test speech enhancement and separation in the wild, as shown in Fig. 4.11d for the multiple speaker case. Although making a real-world system based on multimodal sensing, and end-to-end deep learn-

ing frameworks involve additional challenges due to Lombard effect [170], potential interference, and possible covariate shift in the neural network layers, we try to explore whether there would be improvements compared to an audio-only system. We ask a user to read Rainbow and Arthur passages (details in [48]), and play background noises from a pair of speakers. Since this experiment does not have the ground truth clean signals, we only evaluate the performance in terms of word-error-rate, and character-error-rate. To have a fair comparison, we ask the users to read the same material in another quiet environment and capture the performance in that setting. We use Google's speech-to-text engine without any model adaptation to construct transcripts. As our speakers are not native speakers, and the *RadioSES* is implemented with telephone-quality speech (8 kHz), the overall error rate is higher. On the other hand, as presented in Table 4.8, *RadioSES* can enhance and separate multi-person mixtures and outperform the audio-only baseline for speech separation. We also provide example files on our webpage at `https://zahidozt.github.io/RadioSES/`.

4.7.6 Noisy and Partial Input Data

In this experiment, we corrupt input signals by adding noise and zero-padding, which helps us to gain insight into the performance changes when people are further away, or when there is package loss in the system. These experiments are done with the first 3-seconds of the audio streams, as longer audio streams already require some zero-padding or overlapping block processing.

Noisy data: We add white Gaussian noise to obtain radar data at varying SNRs from 20 to -10 dB levels, and report the performance metrics in Table 4.13a. At larger distances, radio signals



(a) Noisy radio inputs. ∞ represents undistorted radio signals, which actually include device noise.

(b) Shorter radio inputs

Figure 4.13: Performance for distorted radio inputs. Dashed lines represent the performance of the audio-only baseline

are expected to be noisy, and this experiment explores when the radio signals are still useful. *RadioSES* outperforms audio baseline, until a radio SNR of $-5dB$. When the radio signal has further noise, a similar performance as the audio baseline is achieved. This experiment indicates that there is great potential for *RadioSES* at larger distances.

Partial input: In this experiment, we zero pad the radio streams to reduce the available radar stream duration and test input radio durations of 2s, 1.5s, 1s, and 0.5s. Such configurations can be used when there are power requirements or package loss in the radio stream. As shown in Figure 4.13b, *RadioSES* can still help with speech separation tasks and improve the performance, compared to the audio-only baseline, when there is at least 1s of signal (i.e., 33%), in terms of perceptual quality. *RadioSES* system performs better than the audio-only baseline with respect to all inputs after 1.5s of inputs. This indicates that for power-constrained settings, *RadioSES* can be operated with a duty-cycle less than 50%, and can still bring performance improvements, along with the aforementioned benefits of source association.

4.7.7 Partial Detection

Although having speakers outside the FoV of the radar is not our key focus in *RadioSES*, we explore the limits of *RadioSES* in such a mode of operation by allowing one speaker to be outside the FoV. This setup requires the use of alternative approaches to estimate the number of speakers, as the radio-based methods will output fewer people (In practice, we may still use radio-based estimation by leveraging temporal information). We zero pad a radio stream to simulate no information from the outside user, and investigate whether *RadioSES* can benefit from having partial information. We evaluate a single person's missing case, but an extension to two missing people is also possible, with permutation-based methods. As shown in Table 4.9, *RadioSES* can still outperform the audio baseline with a large margin, and improve the performance, with missing people. We do not observe much performance decrease in 2-person noisy mixtures, when one person is outside. For 3-person mixtures, there is more decline, but the gap between audio-only system is larger, and benefits of having the two other radio signals are clear.

4.7.8 Ablation Studies

In this experiment, we train RadioSESNet without several blocks to understand the effect of each component. We use clean 2-person mixtures for our ablation study. As shown in Table

Table 4.9: Performance for partial detection of sources

Case	2-person (noisy)			3-person (noisy)		
Metric	AO	AR(1)	AR(2)	AO	AR(2)	AR(3)
SiSDR	7.7	10.1	11.2	4.9	8.3	9.3
SIR	18.2	20.7	21.0	13.0	17.7	19.2
STOI	0.74	0.81	0.81	0.74	0.81	0.83
PESQ	1.95	2.19	2.20	1.68	1.89	1.96

Table 4.10: Ablation Study

Model	SiSDR
RadioSESNet	15.4
w/o Radio DPRNN	15.2
w/o Any Radio	13.5
w/o Audio DPRNN	4.8
w/o HP filter	0.1

4.10, we remove i) Radio DPRNN blocks ii) Audio DPRNN blocks and iii) High-pass (HP) filter from the mask estimation. In the last case, the audio stream is still used to encode the signal, in order not to change the main structure of *RadioSES*, but is not passed through any DPRNN blocks. As shown, DPRNN blocks of the radio stream bring 0.2 dB overall improvements. Using radio signals alone for mask estimation (w/o Audio DPRNN) still results in a SiSDR of 4.8 dB, showing the usefulness of the radio channel alone for mask estimation. When the high pass filter is removed *RadioSES* cannot operate, as the low-frequency components (dc bands) dominate the processing pipeline.

4.8 Discussion

In this chapter, we propose *RadioSES* to improve the robustness and performance of SES tasks using radio modality. Despite promising results with *RadioSES*, there are certain limitations and many interesting directions to pursue further.

Other side channels: Although in this work we assume the vibration sources in the field-of-view of radio device to be from vocal folds only, radios can also measure vibration of other sources, such as guitars [45], or machinery [3]. These vibration sources usually create some sound signature, and they can be used to estimate the sound from each source separately, as done using cameras in [171].

Microphone arrays: *RadioSES* uses a single microphone along with an mmWave sensing device. On the other hand, it is also possible for *RadioSES* to work with a microphone array and radio modality can still bring further improvements to overall performance. Although beamforming in microphone arrays may indicate that radio modality is unnecessary, it can fail in noisy or reverberant [157] environments. Since *RadioSES* senses the vibration of the *source*, it can estimate the direction of the sound for robust beam-steering or can extract the source vibration without any reverberation for further improvement. Some recent work addresses this problem in audiovisual domain [172], and we believe similar contributions using *RadioSES* can be achieved in the future.

Moving Speakers: Currently, *RadioSES* is designed to track bodies with the assumption that they do not move significantly. This is usually a common constraint in the relevant vital signs monitoring literature (breathing, heart rate), although some recent work started addressing motion for breathing [156]. A more thorough system should support medium and high levels of source motion. To that end, coherent combining of multiple vocal fold bins from person point clouds (*e.g.* [37]), or deep learning [156] can be some interesting ideas to support multiple moving targets.

Sensing Distance: Our experiments indicate that *RadioSES* can work robustly until the speakers are 1m away from the device, and preserve the gains compared to the audio-only baseline. The performance for both cases decreases, which is due to the training *audio* dataset being captured from a short distance. However, the performance improvements from *RadioSES* do not decrease much with the distance. During our experiments, we realized that raw signal SNR is still high at large distances (*e.g.* 2.5m) for people with low pitch (*e.g.* males). To support all users, we limited the practical range to 1m, much larger than the range of using ultrasound [63]. Although

not much radar signature can be captured from these bodies when they are further away, they can still be robustly detected, (*e.g.* as in vital sign monitoring), and even the reduced number of high quality radio streams can still help to improve the performance, as illustrated in Section 4.7.7. Moreover, a different hardware can capture vocal folds vibration from $7m$ in [46], or at $50m$ [173]. We believe *RadioSES* can benefit from better hardware significantly, and a more practical system can be built.

Multipath Effects: In our experiments, we consider cases with multiple sources in front of the radar, and training data assumes perfectly clean radio streams for each person. However, in challenging conditions, wireless sensing-based systems can have a strong multipath effect. Although in mmWave bands, the effect is not as detrimental as 2.4/5 GHz, it can still reduce the performance. We did not encounter this issue in our short-range experiments, but it can be a limiting factor for long-range indoor sensing. We plan to address this issue in the future by potentially simulating multipath data.

Power Consumption and Cost: Although our evaluation board costs $\$300$, a single mmWave device can be purchased for $\$15$ from TI. Transmission power of the device is 12 dBm ($\approx 16mW$) and the selection of radar parameters result in a duty cycle of 7.3%, (*i.e.*$\approx 1.2mW$). For comparison, the size of these devices can go as small as $6mm \times 6mm$ to fit in a phone [16], and the power consumption of the radar in that phone is $1mW$ [16]. Furthermore, *RadioSES* does not require capturing the entire signal duration (Section 4.7.6) and based on the application, lower power consumption can be achieved by reducing the duty cycle further down. As there are already devices with continuous mmWave sensing capabilities, we believe *RadioSES* is feasible to be integrated with smart devices, and this work introduces a new application.

4.9 Summary

In this chapter, we present *RadioSES*, a joint audioradio speech enhancement and separation system using mmWave sensing as a side channel information. *RadioSES* improves the performance of existing audio-only methods with the help of radio modality and achieves similar improvements as audiovisual systems, with further benefits in computation complexity, privacy, and potential NLOS operation features. Furthermore, *RadioSES* can detect the number of sources in the environment, and associate outputs with the physical speaker locations, all being challenging problems in audio-only domain. Real-world experiments show that *RadioSES* outperforms the state-of-the-art methods considerably (*e.g.* 3 dB SiSDR improvements in 2-speaker mixtures w.r.t. audio-only baseline), demonstrating the great potential of audioradio SES.

Chapter 5: *RadioVAD*: Radio-based Voice Activity Detection System

5.1 Introduction

Voice interfaces have become one of the key elements of human-machine interaction in recent years, with the widespread availability of smart assistants. For most voice interfaces, whether a single microphone to record sound or a multi-microphone array to process and understand the user commands, a voice activity detection (VAD) algorithm is the first essential processing block. A robust VAD system enables the removal of silent and unrelated sound segments, prior to transmission and processing, and therefore helps to reduce the computational complexity and power consumption.

A high-performance VAD has many use cases, especially when the voice is transmitted to another human party, such as in meetings and conference calls. Oftentimes, meeting attendees in noisy and interference-prone environments (*e.g.* people in open workspaces with multiple nearby people) need to toggle on and off their microphones manually, as the interference creates discomfort to the party listening on the other side. In a challenging environment, such as teleconferencing while driving, the speaker generally needs to interrupt the device through physical interaction (*e.g.* touch or gesture), which is usually illegal and dangerous. Smart assistants and hands-free systems are also not applicable in such scenarios, as the voice data is already active and transmitted. In other words, the user cannot *ask the smart speaker to turn off the*

microphone, as the microphone is actively transmitting the vocal commands to the other listening party. In a different scenario, a smart speaker may need to be activated only by a particular user and remain deactivated when other interfering speakers are existent. For example, in an open-space environment, the smart speaker of a particular user can be activated by users in other desks. Furthermore, if there are multiple smart assistants within a room, different areas of the room can be assigned for a particular smart device, to prevent multiple smart assistants to trigger simultaneously (*e.g.* [174]). An automated high-performance VAD system that has *spatial* sensing capabilities would bring tremendous advantages to these practical scenarios, as it would minimize the need for user interaction and improve the quality of the voice calls significantly.

In short, an ideal VAD for practical scenarios should have the following properties:

- **Robustness against interference and noise:** As mentioned previously, an ideal VAD system should be robust to arbitrary background noises and interference signals. The system should be able to combat *arbitrary* signals to be a practical system.
- **Computational Complexity:** Since a VAD system usually runs continuously in the background (either with hotword detection or not), it has to be computationally efficient, and has low power consumption.
- **Detection Delay:** To enable practical applications, an ideal VAD should be responsive with minimal processing delays.
- **Spatial Sensing:** In order to focus on the *target* speaker, an ideal VAD system should be able to selectively extract the voice activity of the desired user.

We name a next-generation VAD system with these capabilities with *irVAD*, where the name

stands for *interference and noise resilient voice activity detection system*. To develop an *irVAD*, a system needs to use auxiliary information about the source, as microphones are inherently prone to interference. An *irVAD* usually cannot be implemented by noise and interference cancellation, as these are computationally complex tasks and they do not provide additional information about the *particular* source if there are multiple sources. Some recent work addresses this problem by speaker-conditioned voice activity detection [104] that is only triggered based on the speech of a particular speaker; however, this requires the collection of *a priori* data. Other methods involve conditioning speech on the physical characteristics of the user. As an example, a video [111] of the user can be used to selectively detect voice activity. On the other hand, this method relies on lip motion, which may not be available, especially when the users are wearing face masks. Furthermore, this requires perfect lighting conditions and can potentially raise privacy concerns.

In this chapter, we achieve a breakthrough of *irVAD* by exploring a second modality beyond microphones, mmWave. The motivation to use mmWave is multifold:

- Unlike microphones that capture *ambient* sound, mmWave radars can separate sources in the environment with respect to their distance and angle,
- They can measure a side channel of speech, vocal folds vibration, remotely
- Their large bandwidth and high frequency enable precise localization of vibration sources and therefore spatial sensing.
- They already exist in some smart devices to perform many interesting sensing applications, such as gesture recognition, breathing and sleep monitoring, and are computationally efficient (deployable within mobile devices) devices.

Using mmWave-based sensing, a system can locate the source of vibration (from vocal folds), with high precision, and use this information to extract VAD as an interference-resilient method. As vocal folds generate the excitation signal for *voiced speech*, it is a good indicator of *individual* voice activity. In contrast to other modalities, such as ultrasound, WiFi, and ultrawideband, mmWave can separate sources with high precision. They do not raise privacy concerns as much as cameras, and unlike lidar or infrared, they exist in many devices; making mmWave an ideal candidate for *irVAD*.

In this chapter, we show that an mmWave-based voice-activity detection system can mitigate the aforementioned issues to build an *irVAD* by a *source specific VAD* method. Assuming that the vibration source lies in front of the radar (possibly with or without a specific location, such as the driver seat of the car) an mmWave-based system can extract the voice activity of the speaker and control the microphone automatically. Our contributions are the following:

- We illustrate the usefulness of radio modality for VAD through mmWave sensing by proposing a radio-based VAD system for *irVAD*, and evaluating it by building multiple silent datasets and using speech datasets.
- We evaluate our results in different areas, with physically different locations in unconstrained settings, and provide extensive comparisons with audio-based methods.

5.2 System Design

In this section, we first describe the VAD problem, illustrate the potential of using radio-modality for VAD task, present the system overview, and give further details of *RadioVAD*.

5.2.1 Design Considerations

VAD is a binary classification task based on time series data as input. The input is usually taken as a window with a short duration (*e.g.* 32 ms), and an aggregate decision is made for potentially overlapping windows. An *automatic* VAD should be able to detect the presence of the voice of a particular user with no additional user input and minimal constraints. Minimizing user input eliminates the possibility of assuming *a priori* user data, since this requires training. Therefore, speaker-conditioned VAD systems that rely on speaker embeddings or facial data cannot be a solution. Another potential auxiliary information is source location, and the system can activate according to a particular direction. Although microphone arrays enable filtering sources according to their incident angle with the help of beamforming, they cannot distinguish a *nearby target user* and a background user. Furthermore, beamforming can easily fail in noisy or reverberant conditions [157]. Consequently, the proposed system cannot rely solely on beamforming. To constrain the VAD on the source vibration, capturing the distance (*range*) and incident angle (*azimuth*) of the source with high precision is needed, which is not possible by audio-systems alone without additional further assumptions.

In addition to these, a VAD system is required to be robust, computationally efficient, real-time, and responsive, since it is usually a preprocessing block for many applications.

5.2.2 Feasibility Study

Before explaining how an mmWave-based system can solve the aforementioned limitations, we start by illustrating the feasibility and potential of radio based VAD in Fig. 5.1, where we have a 20-second long audio and radio capture from an environment, where some background

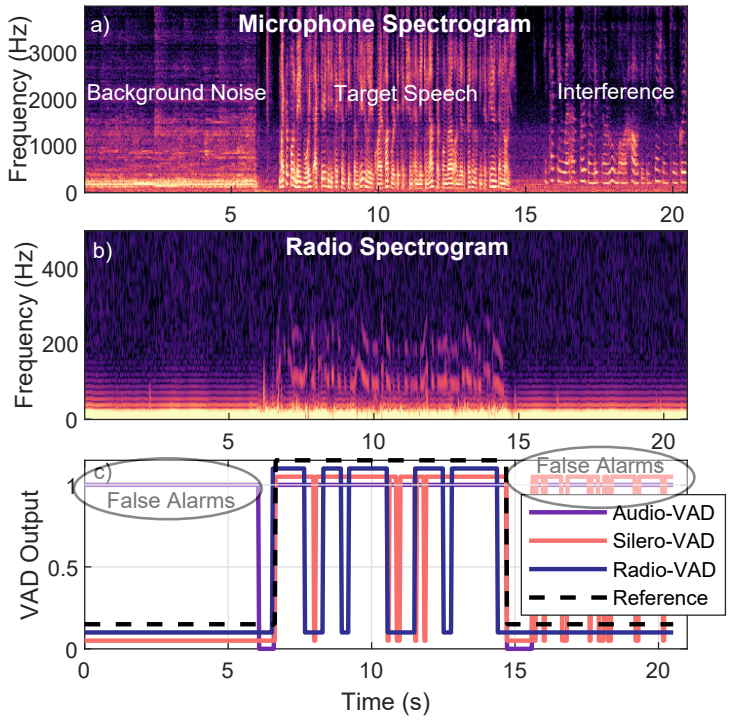


Figure 5.1: Illustration of the Concept. a) Microphone spectrogram, which has noise (0 to 6s), target speaker (6s to 15s), and interference (15s to 20s), b) Radar spectrogram showing activity only during target speech, c) Resulting VAD from trained methods and Silero VAD

noise is played by external speakers, followed by the target and an interfering speaker, respectively. As shown in Fig. 5.1a, the single microphone captures ambient signal without being able to separate different sources. On the other hand, in Fig.5.1b, we see that the radio spectrogram only includes vibration from a particular user and is not affected by background noise or interference. Fig. 5.1c further displays detection results from our trained Audio-VAD, a reference audio VAD (Silero-VAD [175]), and our proposed radio-based VAD. As shown, both audio VADs are triggered when there is interference, whereas a radio-based system is robust against interference and can only be triggered by the target user. We also note that radio-based detection can preserve correct decisions even when there is concurrent talking, which will be discussed later.

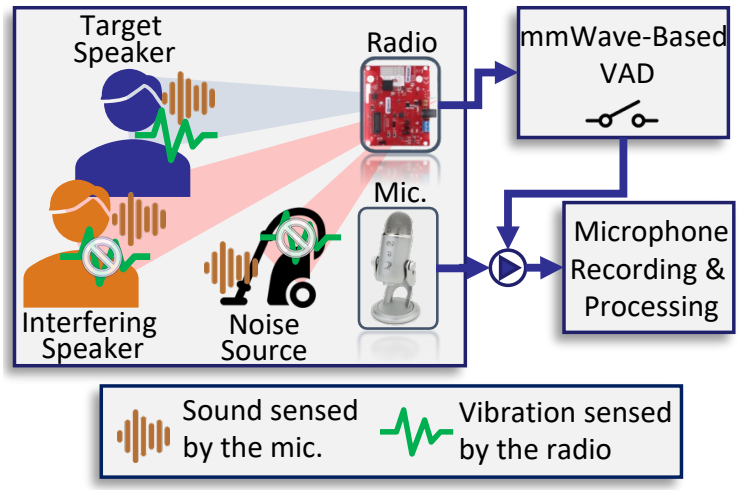


Figure 5.2: System Design. A multi-sensor device (radio and microphone) detects the activity of the *target* user through radio-modality, and activates the microphone for further processing. It is robust against other noise and interfering sources, illustrated in Fig. 5.1

5.2.3 System Overview

In our proposed system, we assume that the smart device has a microphone and an mmWave radar, similar to Google Nest Hub [17]. The main assumption of our system is that the *target* speaker is in the field of view (FoV) of the radar. As illustrated in Fig. 5.2, our system relies on a radio-based VAD alone, which runs continuously in the background (potentially along with other sensing applications), and triggers the microphone recording for further processing. These further tasks can include speech recognition, speech-to-text conversion, and speaker enhancement, all of which benefit from a robust VAD and are natural extensions of this system. To achieve *irVAD*, our system consists of two main blocks. First, we briefly describe the feature extraction block, which is analogous to face detection and tracking methods in audiovisual literature and is needed to ensure speaker conditioning on the *source vibration*. Second, we introduce the neural network structure for our VAD tasks, which is designed to satisfy real-time and low computational complexity requirements.

5.2.4 Feature Extraction

RadioVAD relies on the raw radar signals, which are complex-valued time-series data, similar to *RadioSES*. Based on beamforming and frequency modulated carrier wave (FMCW) technique [121], radars can extract a time-series data from each distance (*i.e.* range bin) based on some granularity (*i.e.* range resolution), and from different angles (*i.e.* azimuth bin) with the resolution depending on the antenna array. We assume that the range-azimuth plane of the radar signal is available through the appropriate radar processing operations, as explained in [121].

Based on the range-azimuth data, the system requires *source detection* and *localization* of the candidate range-azimuth bins. To that extent, we use a variance-based detection scheme to find the *nearest* user to the device, which is a good indicator of presence, due to the body motion caused by breathing and used extensively in the vital-sign monitoring literature for localization [35, 37]. Since this is a widely adopted block in the literature, we refer the reader to the related work and assume that we have a time-series data extracted from the human chest and throat. We have further explained a suitable detection block in Chapter 4 and refer the reader to that chapter.

5.2.5 VAD Network

Like most of the recent VAD works, we evaluate the performance of our system using a neural network (NN), which is depicted in Fig. 5.3. We select the structure of our neural network based on the previous relevant work that explores speech enhancement and separation [84], and is shown to give good performance with stable training, while being compact and near real-time. We use the same NN in [84] with minor changes. The structure of the NN is as follows: First, we obtain a time-frequency representation of the input radio (and also audio)

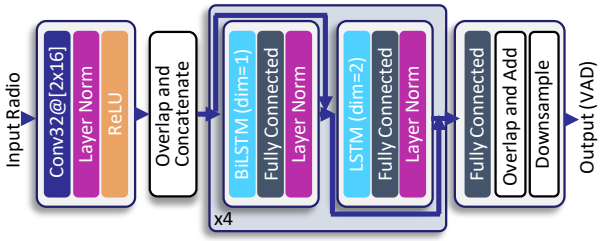


Figure 5.3: Neural Network Structure

signal with a 1D convolutional layer. Through the overlap-and-concatenate operation, we obtain a 4D representation of the input. Later on, the 4D structure is passed through BiLSTM, fully connected and normalization layers. These are followed by another set of LSTM, fully connected and normalization layers, and the same structure is repeated 4 times. All these layers preserve the dimensionality of the input. The output is reshaped to match the input data dimensionality through a fully connected layer and overlap and add method. Last, the output is downsampled to VAD sampling rate through averaging. Further details about DPRNN can be found in [84].

We make comparisons with an audio-only baseline and our proposed system, using the same NN, with a variety of datasets to illustrate the feasibility. When using the audio-only system, the sampling rate of the input increases 8 times, which increases the model size and computational complexity. Therefore, even if the audio and radio models match the same performance, a radio-based system has 8 times lower computational complexity due to the lower sampling rate. In terms of comparison, our radio-based neural network includes 25.8k parameters, which is quite compact. In contrast, our audio-only baseline included 360k parameters, at a sampling rate of 8 kHz. Consequently, achieving similar performance to audio baseline with radio modality indicates a computationally efficient method, and shows great promise of radio modality. Later on, we also investigate a multimodal system to further improve the performance of the system and illustrate the benefits of *RadioVAD*.

5.3 Experiment and Implementation

In this work, we use the dataset explained in Chapter 3, which includes mostly voiced audio and radio files, which had joint audio and radar recordings of 19 users from 5700 sentences. This dataset only includes static users who were allowed to move naturally during speech but not much; therefore, it is not sufficient to evaluate the performance of a VAD. Furthermore, it also lacks silent audio and radar recordings, as the data was cropped with respect to the beginning and end of the sentences.

To overcome these limitations, we collect additional data from 11 participants in our experiment area (Loc I), which can be seen in Fig. 5.4. We ask the users to sit in the designated area, where they are approximately 0.5m away from the radar. The users work freely in the environment with no further instructions on how they work, except to remain silent. Their work routine included using a separate laptop, and the provided monitor, reading from/writing to paper documents, checking mobile phones, and typing on the keyboard. Each user was asked to work for an hour in the given location. After subtracting the overhead from the data capture, the data from each user ended up around 35 minutes, with the total data being around 6 hours. Furthermore, we have collected additional radar and audio data in other locations to improve robustness. One dataset includes 30 minutes of new location data to test further generalizability of the system. In addition to these, we further collect data in more challenging scenarios, as will be described later. Some of these challenging scenarios are driving, moving the device intentionally, and making other motions with mouth, such as whispering or gumming.

Data Labeling: In order to generate reference labels, we use a high performance off-the-shelf VAD on clean audio files. We extract raw detection decisions from Silero VAD [175] with 32ms-

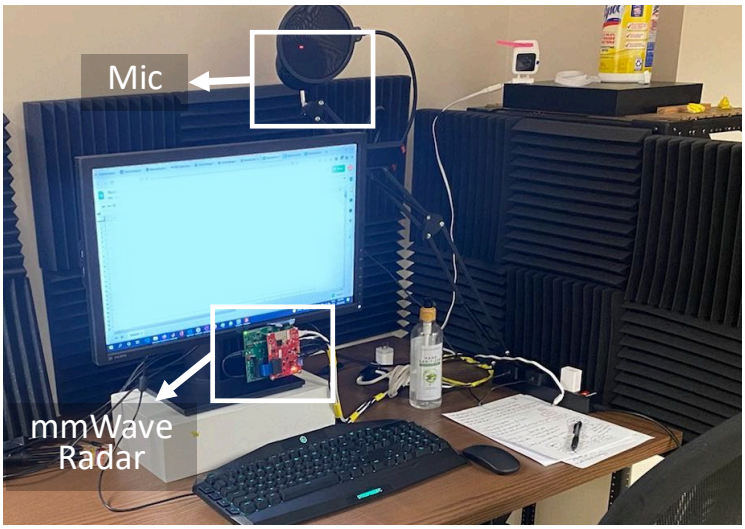


Figure 5.4: Data Collection Setting

long decision windows and smooth the decisions by setting a minimum speech duration of 0.25s and a minimum silence duration of 0.1s. We set onset and offset thresholds as 0.5 and 0.35, and process the data in a causal fashion. As an example, a detection is preserved, if there is no silence within the first 0.25s, and a nondetection is kept if there is an uninterrupted silence for 0.1s. In our quiet data setting, we set all reference labels to zero, as we asked the users to be silent. To train the audio-only system with background noise, we create the same speech enhancement dataset in [52], which combines the clean audio files with the noise files in WHAM [145] dataset. Furthermore, we also use the noise files in WHAM with the decisions being 0. Having both noise files and clean audio files corrupted with the noise allows us to mitigate overfitting issues, as the NN can easily learn to distinguish *the environment* in a different scenario. For evaluating the silence case, we use a subset of data collected in location (Loc.) I (Silent Set I), and we use all of the data in Loc. II (Silent Set II). Loc. II has never been seen by the NN, whereas some data from Loc. I has been used for training.

Radio processing: We use a variety of preprocessing methods introduced in [53, 156], such as

high-pass filtering and random phase rotation of complex-valued signals. Although our decision windows are 32ms long, we use longer duration samples to exploit *contextual information*. We choose to use 320ms long audio and radio data streams to train our NN, since the dataset in [53] usually includes a voice starting at 0.2 to 0.6s. Therefore, using a much longer window (*e.g.* 3 seconds) can give unrealistic results, as the *starting point* of the speech would always be at the beginning of the long window. During our training, we pick a 960ms long segment from the audio and radio files randomly to ensure a random starting point for each sample. We have also experimented with 320ms long segments but did not observe much difference in performance; hence, we do not report those.

Training: Our training procedure uses a modified F1 score (F_β [176]) loss between our reference and estimated values. F_β score is a modified F1 score, used to balance the cost of precision and recall rates, and is given as:

$$F_\beta = \frac{(1+\beta^2)\text{TP}}{(1+\beta^2)\text{TP} + \beta^2\text{FN} + \text{FP}}, \tag{5.1}$$

where TP, FN, and, FP denote true positive, false negative, and false positive rates respectively. We choose the β as 0.35 to offset the class imbalance between voiced and unvoiced data segments. We have experimentally verified the optimality of this choice, in contrast to the F1 score, and $\beta = 0.5$ and $\beta = 0.25$.

For training, we always use different users in the training and test set for both datasets, but we also provide the performance metrics for the users in the training set (closed condition) to better understand the generalization performance. Since the F1 score is an aggregate metric, we also provide additional evaluation metrics in our experiments.

Implementation: Our NN model and training process have been implemented in PyTorch. We use adam optimizer is selected, with a starting rate of $5 \cdot 10^{-4}$, and we halve the learning rate if there is no improvement on the loss function for 10 consecutive epochs. We implement an early stopping criterion of 25 epochs and train each system for 200 epochs. We use a batch size of 48 samples. Our training hardware includes an NVIDIA RTX 2080S.

5.4 Evalution

In this section, we evaluate the performance of *RadioVAD* in different scenarios, with respect to a variety of metrics and experiments. We first explain the evaluation metrics in Section 5.4.1. Afterward, we present the overall performance of *RadioVAD* in Section 5.4.2, with evaluation of false alarms in a variety of daily scenarios in Section 5.4.3. In Section 5.4.4, we investigate the effect of motion interference and a variety of noise sources. We follow up with a comparison of two modalities and a multimodal system in Section 5.4.7.

5.4.1 Evaluation Metrics

We evaluate the performance with respect to the metrics, such as accuracy, precision, recall, F1-score, and area under curve (AUC). Some of these metrics are given as follows:

$$\text{Precision} = \frac{\text{TP}}{TP + FP}, \tag{5.2}$$

$$\text{Recall} = \frac{\text{TP}}{TP + FN}, \tag{5.3}$$

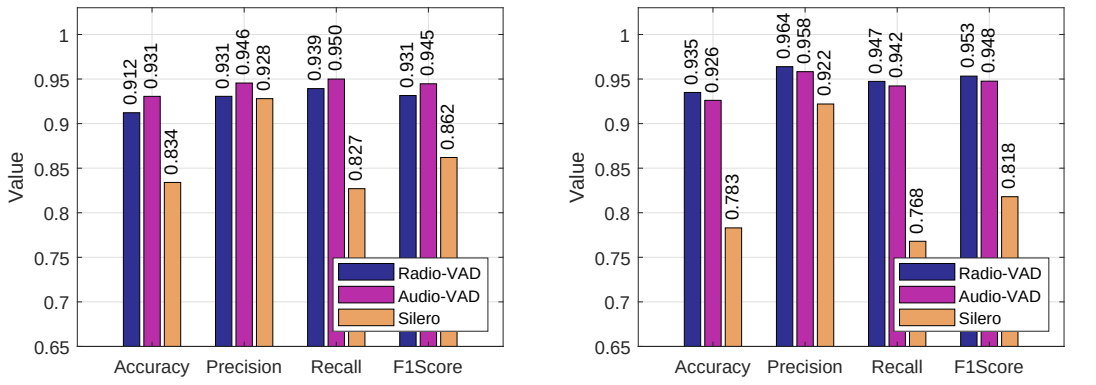
$$\text{F1-Score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5.4}$$

We further evaluate the performance with respect to a variety of environmental factors, such as distance, orientation, occlusions, and arbitrary motion. Moreover, we provide important metrics, such as detection delay, and investigate the effect of user diversity.

5.4.2 Overall Performance

We present the performance metrics in two test cases in Fig. 5.5, both of which are from the speech dataset in [53]. In addition to our *RadioVAD*, and Audio-VAD baselines, we also use an off-the-shelf VAD detector, Silero VAD [175] in this setting to provide a baseline. We note that, Silero VAD is quite robust VAD against noise and has been trained with much larger datasets. On the other hand, our audio baseline is trained in a single (or in a few location) data; and may have some overfitting issues to the background In the first test set, we evaluate the unseen text from the users used during training, whereas test set II only constitutes unseen users. First, we observe that even though the radar captures secondary information, it can still match the performance of the audio-only method, in terms of accuracy and F1-score closely in the test set I. In test set II, the proposed system outperforms the audio-VAD, showing the promise of radio-based system. In addition, *RadioVAD* provides much higher performance than Silero VAD, and gives very similar performance to the audio baseline. We note that our audio training pipeline only includes noise files from a particular dataset (*e.g.* WHAM), and its performance can potentially decrease with a wider variety of noise files. Silero VAD is a very good VAD in practical scenarios and the performance gap between *RadioVAD* and Silero VAD is clear. In addition, we observe that *RadioVAD* outperforms the audio baseline in the unseen condition which is an indicator of the generalization capabilities of the radio-based VAD method. In summary, using the side-channel

information, an mmWave-based system can match and surpass the performance of a microphone based system. We investigate *at what conditions*, *RadioVAD* is better than an audio based system in the next sections, and change our focus to detection delay.



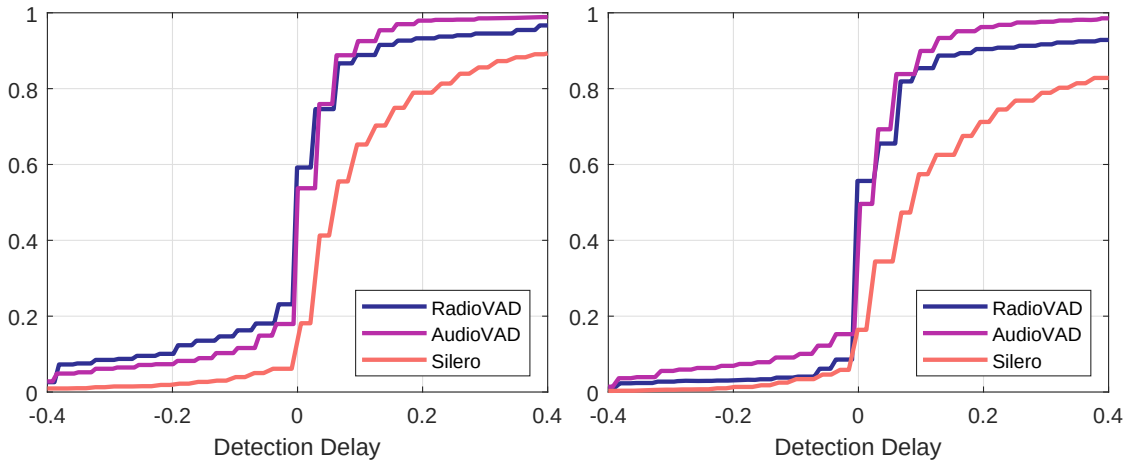
(a) Performance in Test Set I (Closed Condition-Seen Users)

(b) Test Set II (Open Condition-Unseen Users)

Figure 5.5: Performance Comparison of *RadioVAD* with Audio VAD and Silero

Detection Delay: For a high-performance VAD system, detection delay is of utmost importance, as this will trigger the capturing of audio signals. Consequently, we evaluate the performance of *RadioVAD* with respect to the detection delay and provide scatterplots and cumulative distribution functions of the delay in Fig. 5.6. As can be seen, more than 85% of the detections of *RadioVAD* have a delay less than 64ms, which can be acceptable, and the median detection delay is 0ms. In addition, *RadioVAD* outperforms Silero VAD, and matches the performance of the audio baseline for most of the time. We realize that some motion prior to speech helps *RadioVAD* to make early detections (*e.g.* inhaling), which should be acceptable for the scenarios of interest in this work.

5.4.3 False Alarm Performance



(a) Test Set I (b) Test Set II

Figure 5.6: CDF of detection delay for *RadioVAD*, audio baseline, and Silero VAD

5.4.3.1 Natural Working

Third, we test our methods in the silent data setting and present the results in Table 5.1. In this figure, Loc 2 has been used for silent data collection, and all of the data in this location has been silent (except for natural sounds during working). On the other hand, Location 3 has never been used for training and is in a more active environment with nearby coworkers. When the silent data is captured in a single location, the neural network can potentially learn the background sound signature of that location and overfit, which is indicated by huge performance loss in Location 3 for our audio baseline. In all settings, the radio-based method works consistently and results in a similar amount of false alarm rates. We also note that, the performance of *RadioVAD* matches that of the Silero-VAD, which is shown to have extremely good performance. Last, we note that these experiments only include background noise, without any interference (person speaking nearby), and the benefits of the radio channel are even more prominent in those cases, as illustrated in Section 5.2.2.

Table 5.1: False alarm rate comparison in silent settings

Method	Silent Set I (Location II)	Silent Set II (Location III)
RadioVAD	4.83 %	**4.78 %**
AudioVAD	**0.19 %**	16.84 %
Silero	3.90 %	9.91 %

5.4.3.2 Arbitrary Motion

In this evaluation, we conduct additional experiments to test the false alarm rate of *RadioVAD* during a variety of motion types. As noted in the previous section, when there is no motion, the radar signals are relatively stable and do not create any signature to be detected. On the other hand, when there is motion, it can potentially trigger the *RadioVAD*. To further validate robustness against the motion, we conduct the following experiments:

- **Eating:** We asked the users to have their lunch in front of the radar at their desks, and not speak while eating. We have forced the ground truth labels to be zero, and measured the false alarm rate due to eating, in *RadioVAD*, Silero VAD, and our audio baseline.

- **Drinking:** Drinking is also another natural movement, which can occur while the radar is operating in front of the user. We asked the users to drink some water and labeled the reference signals with silence.

- **Gumming:** Many people gum while they are working, and a robust system should not give false alarms. We asked users to gum in front of the radar.

- **Silent Speech:** Last, we asked the users to whisper in front of the radar. This involves motion of the lips, and can potentially trigger false alarms.

We have presented the false alarm rates in each of these experiments in Fig.5.2. As shown, except

Table 5.2: False alarm rate comparison in silent settings

Method	*RadioVAD*	Audio VAD	Silero
Eating	11.48%	2.44%	3.34 %
Gumming	1.62 %	2.03%	0.83 %
Drinking	2.47 %	1.89%	0.5 %
Silent Speech	16.1 %	2.53%	2.65 %

the eating and silent speech cases, the false alarm rate is low (less than $\%3$), whereas eating and silent speech introduce some false alarms. We argue that *RadioVAD* uses lip motion to some extent, and some of these motions include trigger false alarms due to the opening of the mouth (*e.g.* silent speech or eating). The overall false alarm trend of *RadioVAD* is comparable with that of Silero VAD. We note that, *RadioVAD* is never trained against these types of motions, and the performance can be improved by adding these corner case data into the learning set.

5.4.4 Effect of Interference and Noise on Radio Signals

As discussed previously, radio signals are inherently robust against acoustic noise and other interferences. Any source that is located away from the *target* source creates minimal to no interference, due to the spatial separation property of radio-based VAD[1]. On the other hand, additional signals at the same *target* location can create interference, and potentially reduce the performance of radio-based VAD. To that extent, we investigate different corrupted radio signals to further verify the performance of the *RadioVAD*.

[1]In rich scatter environments, nearby objects can also create interference, due to the multipath effect. We discuss the effect of multipath later on and assume that the radio signals are not corrupted from the nearby objects, as the multipath effect is rather weak at short distances.

5.4.4.1 Interference

Since radio signals capture the motion (*i.e.* displacement) of objects in the environment, they are also affected by the relative motion between the device and the sources. In other words, if the device or the user moves in the environment, this would introduce Doppler shifts. In order to test the effect of motion signature, we conduct the following experiments:

- **Holding the device in hand:** In this experiment, we asked the user to hold the radar device in hand, and speak naturally. We did not ask the user to make any arbitrary motion, and the device was relatively stable.

- **Moving the device in hand:** In this experiment, the user moved the device intentionally in different directions. The total range of motion was around 30cm, with 8 directions. We asked the user to not change the orientation of the device.

- **Holding a paper in front of device:** In this experiment, we asked a user to read material from a piece of paper. The user held the paper in his/her hand and blocked the line-of-sight from the radar device. We asked the users to read the material naturally.

- **Speaking with hand gestures:** In this experiment, we asked the users to move their hands around their chest region excessively. Both hands moved the entire time, while the person remained stationary.

- **Moving the body:** In this experiment, we asked the user to move his body in 6 directions, up and down, left to right, back and forth.

Table 5.3: Performance in interference cases

Experiment	Accuracy	F1-Score
Holding in hand	97.04%	96.94%
Moving in hand	89.79%	88.23%
Holding paper in between	95.06%	94.43%
Speaking with Gestures	71.26 %	76.27%
Moving body (Left to Right)	94.61 %	95.96%
Moving body (Up & Down)	95.14 %	96.41%
Moving body (Back & Forth)	92.35 %	94.10%
Wearing a face mask	94.76%	94.53%

- **Occlusion:** In this experiment, we asked the users to wear a facial mask when they are speaking.

All of these experiments are conducted in quiet environments to successfully extract the reference ground truth signals using an off-the-shelf detector. Consequently, we compare the performance metrics of these experiments with the static case to investigate the performance degradation. As can be seen in Table. 5.3, the *source* or *target* motion affects the performance of *RadioVAD* minimally. Although these types of motion introduce *interference*, they are superimposed with the source, and the Doppler shift due to the body and source motion is comparatively small. Second, holding a paper in between still allows to capture the radar signals with minimal interference, and does not show any performance reduction. The performance decreases when there is speaking with hand gestures. This is expected, since the Doppler shifts due to the quick hand gestures overlap with the vocal folds frequencies, hence reducing the performance.

5.4.4.2 Effect of Noise on Radio Signals

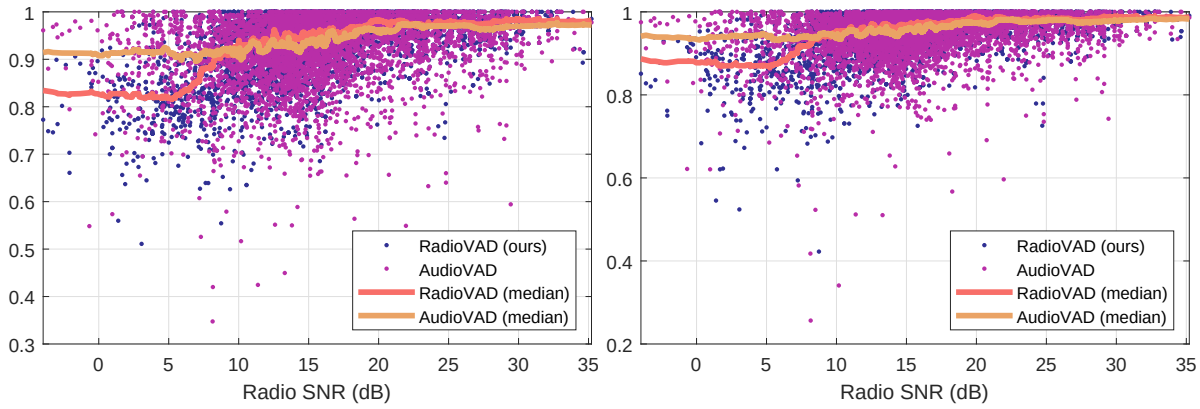
In this section, we investigate the effect of the signal-to-noise ratio on radio signals to better understand the limitations of *RadioVAD* in terms of distance and noise robustness. Our evaluation

is by inspecting the performance metrics with respect to an estimated radio SNR.

In this evaluation, we only use the static recordings, as it is not straightforward to separate the motion signal strength from the vocal folds strength using the radio signals. Radio signals are the superimposition of both motion and vibration signals and measuring the energy in vocal folds bands when stationary is one of the simplest and accurate methods for estimating vocal folds vibration. Even though natural speaking introduces some motion in body and lips, we ensure that most of the signal content is due to speech, by using static recordings. We use the data segments that are labeled with speech to estimate average signal energy, and silent segments to estimate the background radio noise. As shown in Fig. 5.7, the performance of *RadioVAD* increases with the higher radio SNR values. The *RadioVAD* starts to outperform audio-based VAD when the radio SNR is higher than $\tilde{8}$dB. We observe that the precision rates are quite consistent over different SNR values, which is in line with the definition of the precision, *i.e.*, the amount of false positives rely on noise energy, which is usually consistent over different values. Furthermore, we observe a weak correlation between the performance of audio based performance and Radio SNR, which should be indicative of the speaking strength, but the relationship is very minimal. We conclude that, if Radio SNR is greater than 8 dB, *RadioVAD* performs better than audio VAD, along with its aforementioned computational benefits. When the underlying dataset for audio signals have more noise, *RadioVAD* will be more preferable at lower radio SNR points.

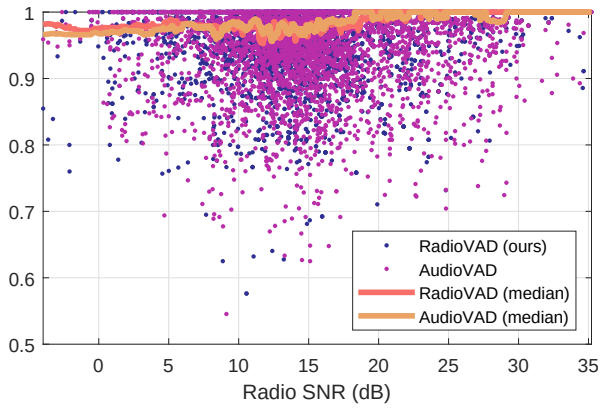
5.4.5 User diversity

Since our dataset is relatively small compared to those in the speech domain, the effect of user diversity needs to be investigated carefully. To make our evaluation more comprehensive,

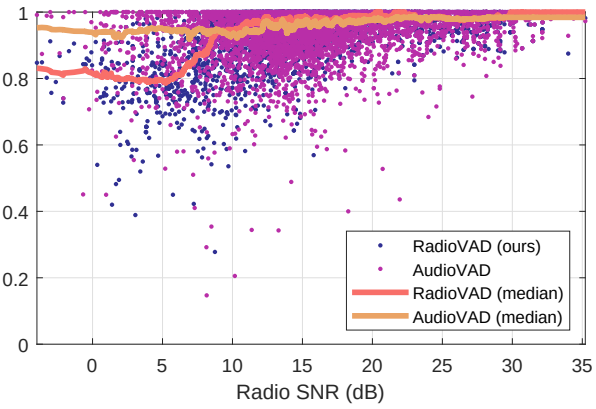


(a) Accuracy w.r.t radio SNR

(b) F1 Score w.r.t. radio SNR



(c) Precision w.r.t. radio SNR



(d) Recall w.r.t. radio SNR

Figure 5.7: Performance w.r.t. SNR

we apply 5-fold cross-validation to our dataset, based on user identities. In each fold, we ensure that at least one of the female speakers is in the testing set, and each fold includes 3 to 4 users to be tested, with no overlaps. We present box plots of F1-scores with respect to different users in Fig. 5.8. We note that, users 2, 5, 10, 11, 13, 14, 18 are female speakers, whereas the rest are male speakers. As can be seen from the figure, *RadioVAD* preserves consistent results between different users and generalizes quite well to all users. Some users have relatively lower F1 scores, which is due to the lower recall rates for these users. We note that only users 6 and 13 show relatively lower performance, whereas the rest are quite consistent. Since these are both male and female speakers, the performance degradation is not due to gender. To better

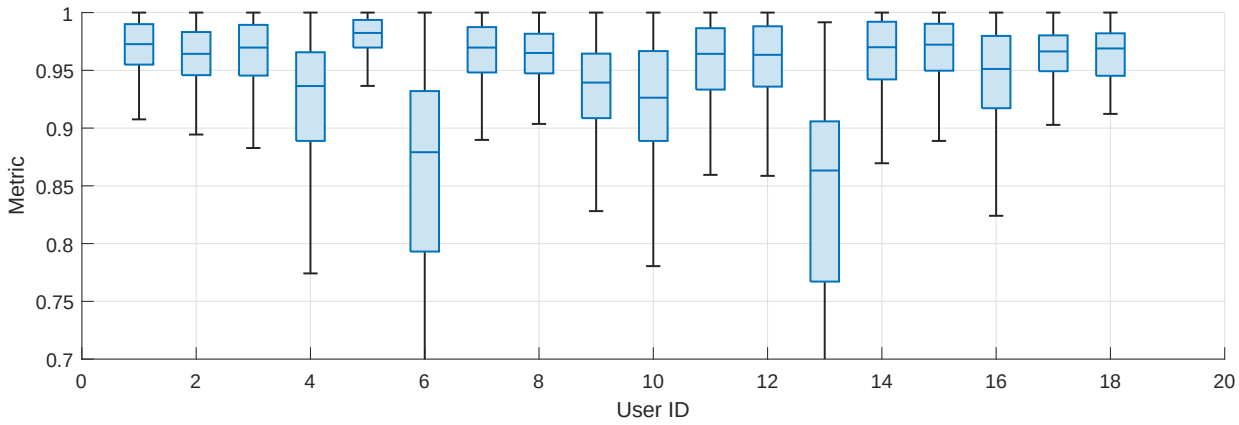


Figure 5.8: F1-Score w.r.t. User ID

understand the performance of *RadioVAD* with respect to the user identity, we provide the box plots of radio signal SNR 5.9. As can be seen, the radio SNR is quite consistent with respect to a variety of users. On the other hand, some users show more variance, such as user 13. This is due to SNR estimation relying on the energy of the radio spectrograms, which also involve motion. Our manual inspection reveals that these users perform more motion in the controlled setting while they are speaking. Among a total of 18 users, 16 users have median and mean radio SNR greater than 8 dB, which is the equal performance point for radio and audio signals, as illustrated in Section 5.4.4.2. Furthermore, we have also noted a higher amount of motion for user 6 in the silent dataset, which involved completely negative classes for the same particular user. Since the neural network can potentially learn the identity of the user, we believe this is due to the training dataset involving more motion from the particular user and can be mitigated by adversarial training methods or creating more diverse datasets.

5.4.6 Environmental Factors

In this section, we evaluate the performance of *RadioVAD* with respect to multiple environmental factors. These include testing the system against changes in distance (Section 5.4.6.1),

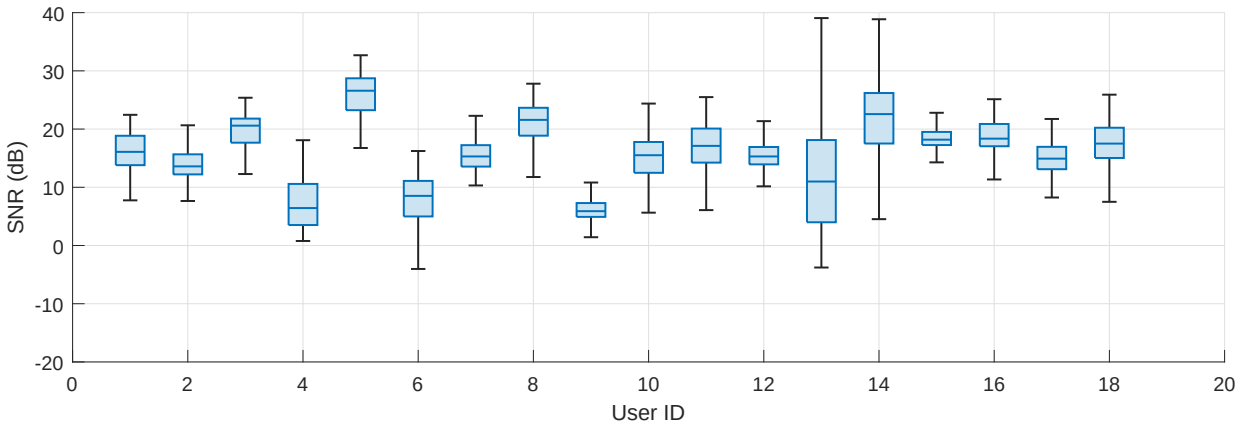
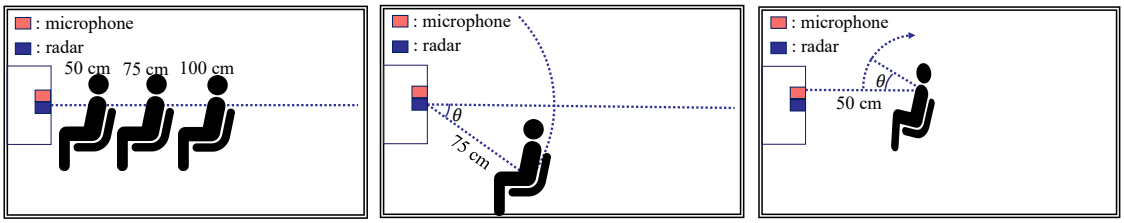


Figure 5.9: Radio Signal SNR w.r.t. User ID

orientation (Section 5.4.6.2), and face orientation (Section 5.4.6.3). Our first three experiments are the same as those in Chapter 4, which are provided here for reference in Fig. 5.10.



(a) Distance setting (b) Incident angle setting (c) Head orientation setting

Figure 5.10: Multiple experimental settings

5.4.6.1 Distance

Most of our experiments are in a controlled setting at a particular location. Operational range is one of the most important characteristics of the practicality of *RadioVAD*. In that sense, we evaluate the performance of *RadioVAD* at varying distances. We only test our system against the ground truth, and report the numbers in Table 5.4. As can be seen, *RadioVAD* preserves its performance until 75cm, with some performance decrease afterward. We believe, this limitation is acceptable for the aforementioned scenarios (*e.g.* driver of the car, person working in front of

Table 5.4: Performance with respect to distance

Exp	Distance		
Case	50 cm	75 cm	100 cm
Accuracy	90.94%	93.03%	83.66%
Precision	91.04%	95.26%	95.74%
Recall	96.13%	94.20%	78.74%
F1-Score	93.47%	94.69%	84.72%

Table 5.5: Performance with respect to orientation

Exp	Orientation			
Case	0°	15°	30°	45°
Accuracy	93.03%	81.85%	92.58%	93.02%
Precision	95.26%	87.46%	95.81%	94.35%
Recall	94.20%	85.04%	93.14%	95.56%
F1-Score	94.69%	86.13%	94.42%	94.95%

his/her desktop, or handheld device), and it can be further improved by better hardware.

5.4.6.2 Orientation

Second, we experiment with the orientation of the human body in the environment. Following a similar set of experiments in Fig. 5.10, we measure the performance of *RadioVAD* at different locations. As can be seen in Table 5.5, *RadioVAD* performs similarly at varying angles. Having a wide field of view is important, and *RadioVAD* can operate at 45 degree angle without a significant performance reduction.

5.4.6.3 Face Orientation

In a practical scenario, the users do not necessarily look toward the radio device, and they may rotate their heads to look around. As an example, a driver of the car can potentially check the mirrors, or a user can look around a screen to see different materials. Therefore, we test

Table 5.6: Performance with respect to face orientation

Exp	Orientation		
Case	0°	15°	30°
Accuracy	92.76%	93.61%	92.42%
Precision	95.13%	93.43%	92.63%
Recall	94.17%	97.26%	96.55%
F1-Score	94.61%	95.27%	94.53%

RadioVAD against head rotation in Table 5.6. As shown, small head rotations up to 30 degrees do not affect the performance of *RadioVAD* and enable high-performance VAD.

5.4.7 Multimodal Systems and Comparison of Audio and Radio

A natural extension of *RadioVAD* is using the two modalities, by proposing an audioradio framework. Although the proposed system can mitigate the aforementioned limitations of the microphones, certain systems may require higher performance. Especially when the computational complexity and power requirements are less strict, an audioradio system can be feasible. To explore the performance of such system, we use the same neural network model, but concatenate the radio and audio channels after the encoder layers. At the output, we map the output directionality to match the dimensionality of the input audio stream and decode the signal accordingly. A high level processing outline is given in Fig. 5.11 model.

In this setting, we provide and compare the performance metrics of the multimodal system with the proposed system in 5.7. As shown, an audioradio model further improves the accuracy, precision, recall, and F1 scores. We note that the audio files are corrupted with the same noise files at the corresponding SNR levels in these experiments to keep the evaluation more consistent.

In addition, a complementary analysis to that in Section 5.4.4.2, we try to answer *at what audio SNR, is it preferable to use Radio-based VAD?*. As shown, when we have corrupted audio

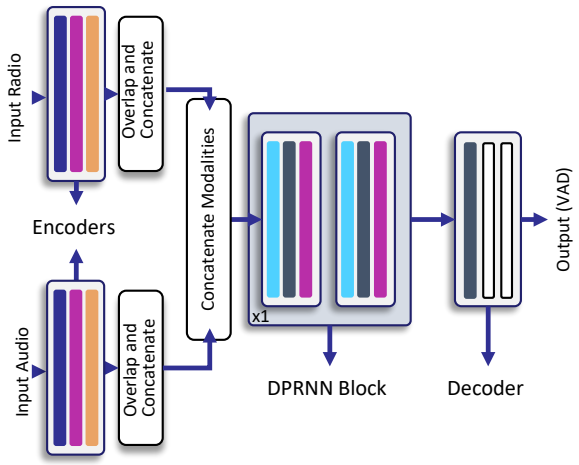
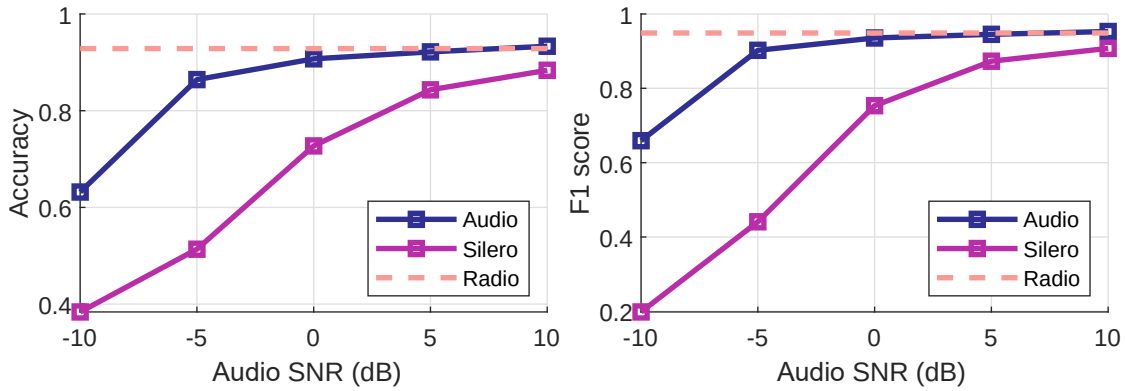


Figure 5.11: Audioradio model, color codes are the same as Fig. 5.3

Table 5.7: Accuracy of multimodal system in different experiments

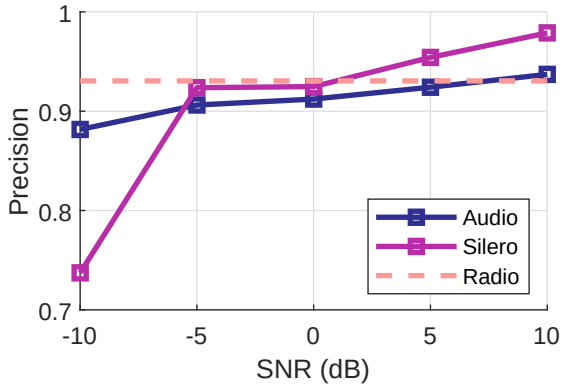
Experiment	CC (Loc1)	OC (Loc1)	Loc2	Loc3	Interference Exp.	False Alarm Exp.
Silero	83.42%	78.43%	96.10%	90.99%	98.85%	98.14%
Radio Only	90.46%	91.83%	98.58%	98.47%	90.34 %	96.84%
Multimodal	90.71%	93.86%	99.04%	98.72%	91.92%	98.99%

signals with varying SNR levels, the performance of *RadioVAD* matches that of audio-baselines, and surpasses them in some other cases (*e.g.* Silero VAD [175]). In order to understand this phenomenon better, we use a predefined SNR value for audio signals and extract the performance metrics. Namely, we use an audio SNR of -10 dB to +10 dB, and run our audio baseline and Silero VAD. We compare the accuracy results in Fig. 5.12. In all cases, *RadioVAD* performs better than both audio-based approaches when the audio SNR is at 0 dB or lower. On the other hand, recall rates, and f1 score matches that of the audio baseline at 5dB, and accuracy is comparable at 10 dB. Consequently, we argue that, whenever the audio SNR is lower than 5dB, it becomes more advantageous to use a radio-based VAD system. This is assuming the radio signal SNR distribution is the same as in our dataset, and the matching point can be even higher (*e.g.* radio at higher SNR can match the performance of audio SNR at 10dB or more).

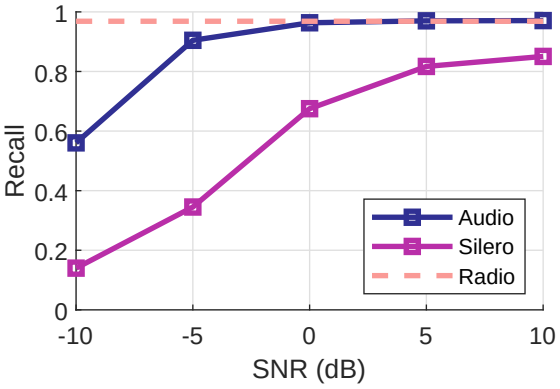


(a) Comparison of Accuracy

(b) Comparison of F1-Score



(c) Comparison of Precision



(d) Comparison of Recall

Figure 5.12: Performance Comparison of Audio-VAD, Silero and *RadioVAD* at varying audio SNR

5.5 Discussion and Future Work

In this work, we evaluate the feasibility of using a radio-based approach for *irVAD*. Our results indicate and demonstrate great potential for radio-based *irVAD*. We also believe that radio modality brings exciting research opportunities beyond VAD to next-generation voice systems. Furthermore, some additional case studies and experimentation is needed to verify and validate the desired properties of the system. These include:

Real-Time Properties: Although our proposed method and deep learning modules run causally, the system includes potential delays in several processing blocks, such as preprocessing, time-

frequency representation, and frame-based processing. An *irVAD* requires a minimal delay in order to enable automatic and responsive triggering of the microphone. To that end, it can also be combined with an audio buffer and time compression to mitigate potential issues. As an example, the system can send several samples before the triggering of VAD by using the audio buffer, which is usually included in smart assistants. When transmitting, the system would apply time compression to send those extra samples within a shorter period.

Motion Compensation: Although our proposed method results in a much lower rate of false positives when there is only motion, we need to evaluate our system in more challenging conditions, such as when there is excessive motion and speech at the same time. Even though the speech (or *vibration*) signals have a distinctive shape compared to the motion, the performance of the system may reduce in such cases. In extreme cases, we believe our system can be combined with a microphone-based system (multimodal) to further improve the performance, as illustrated before.

Multiple Users: In a more relaxed scenario, there can be multiple people in the environment, and they can be speaking simultaneously. Our current system has the potential to detect each person (using body motion-based detector), and make a VAD decision. This means the system can detect the voice activity of each user individually. Our system assumes that the *target* speaker is in a particular location (*e.g.* the closest user). We think that this is a practical assumption, compared to the assumptions in the existing literature (*e.g.* having the face image or speaker embeddings of the user). On the other hand, this assumption can further be relaxed by extracting the radar features of the target user (*e.g.* pitch) and selecting the matching user. In this case, using radar is still advantageous compared to microphones, as the radar signature from the source is not corrupted. Consequently, the radar *features* of the particular user can be extracted even when

multiple speakers are present.

Further Directions: Our system focuses on building an *irVAD* system for higher-order tasks. These include a speech enhancement & separation system, that can be triggered and focused on a particular user, a speaker diarization system that labels VAD of each person separately, or a speech-to-text engine for authorized users or users in authorized spaces (*e.g.* users in FoV of radar). *RadioVAD* can be combined with the *RadioSES* for higher performance as well.

5.6 Summary

In this chapter, we explore *RadioVAD*, an mmWave-based interference-resilient voice activity detector that can be focused on the *sound source vibration*. Thanks to the spatial separation capabilities of mmWave-based sensing, a voice activity detector that is robust against interference from other sound sources can be built. Our extensive experiments indicate great potential for using mmWave for voice-activity detection with the inherent benefits of mmWave such as low computational complexity, privacy preservation, and occlusion resistance. We illustrate that *RadioVAD* can combat interference sources significantly, and match the performance of a microphone based VAD. Our experiments portray the audio and radio tradeoff and show great promise for exciting applications.

Chapter 6: Conclusions and Future Work

6.1 Conclusions

In this dissertation, we have presented the preliminaries of sensing sound and vibration events using mmWave radios, and a variety of systems that include mmWave-based sensing to mitigate limitations of microphone based sensing systems. Our proposed systems aim at enabling new applications using mmWave-based radio devices, in different modes of operations. We explored the potential of using mmWave in speech, sound, and vibration sensing alone in Chapter 2 and 3, and explored its combinations with a microphone in Chapter 4 and Chapter 5.

These include the following:

1. *RadioMic*: mmWave-based Sound Sensing System: In Chapter 3 we propose *RadioMic*, an mmWave radar-based sound sensing system, which can reconstruct sound from sound sources and passive objects in the environment. *RadioMic* is robust to environmental changes, such as those in lighting, and can operate in dark and NLOS settings. Using the tiny vibrations that occur on the object surfaces due to ambient sound, *RadioMic* can detect and recover sound as well as identify sound sources using a novel radio acoustics model and neural network. The flexible design of *RadioMic* enables additional applications, such as sound source localization and classification with unprecedented accuracy,

separation of sound sources with respect to *distance*, and potentially increasing the sound sensing range to multiple rooms. Extensive experiments in various settings show that *RadioMic* outperforms existing approaches significantly and enables many new applications.

2. *RadioSES*: Radio-based multimodal Speech Enhancement and Separation System. In Chapter 4, we introduce *RadioSES*, a joint audioradio speech enhancement and separation system using mmWave sensing. It improves the performance of existing audio-only methods with the help of radio modality and achieves similar improvements as audiovisual systems, with further benefits in computation complexity and privacy. Furthermore, *RadioSES* can detect the number of sources in the environment, and associate outputs with the physical speaker locations, all being challenging problems in audio-only domain. Real-world experiments show that *RadioSES* outperforms the state-of-the-art methods considerably (*e.g.* 3 dB SiSDR improvements in 2-speaker mixtures w.r.t. audio-only baseline), demonstrating the great potential of audioradio SES.

3. *RadioVAD*: Radio-based Voice Activity Detector. In Chapter 5, we explored an mmWave-based interference-resilient voice activity detector that can be focused on the *sound source vibration*. Thanks to the spatial separation capabilities of mmWave-based sensing, a voice activity detector that is robust against interference from other sound sources can be built. Our experiments indicate great potential for using mmWave for voice-activity detection with the inherent benefits of mmWave, such as low computational complexity, privacy preservation, and occlusion resistance.

6.2 Future Works

In this thesis, we have presented practical, real-world sensing systems with extensive evaluations in many cases. On the other hand, to implement these systems in real-time, with more practicality constraints, there is a need to pursue certain directions. These are discussed in the following sections.

6.2.1 *RadioMic*: mmWave-based Sound Sensing System

RadioMic: Our system in Chapter 3 focuses on building a radio-based sound sensing system, that solely relies on surface vibrations. Although *RadioMic* illustrates a great potential for radio based vibration and sound sensing, there are a multitude of directions that can be pursued to make the system perform better, reduce the computational complexity and power consumption, and enable real-time applications.

There are several additional system properties that can be improved further to make the system more practical. In addition, better radar hardware can enable further performance improvements. To summarize, we believe the following are important future directions to make the system more practical.

1. **Training routine and deep learning model:** An integral part of our work was using a deep learning system, that included completely synthetic data, whose performance is reliant on the performance of mimicking the actual data. Although this approach can successfully generate data similar to real-world, and helps to boost the dataset size considerably, it results in lower performance. The overall performance can be improved by using more

realistic data collection and optimal training modules.

2. **Different Applications:** Since *RadioMic* can sense the arbitrary vibrations in the environment, it can be used as a more generic sensing device, and enable additional applications, such as glass-breaking monitoring, fire alarm monitoring, and so on. Furthermore, speech recognition algorithms can be implemented on top of *RadioMic* to further enable voice commands. To improve the performance further, some dictionary constraints can be added to the system. Last, we believe there is a natural extension of the *RadioMic* to detect live and inanimate sources, and a more diverse set of applications can be explored.

3. **Spatially Enabled Sound Sensing:** As using radio signals enables sensing the sound vibration from its *source*, and have its directionality (*i.e.* can only sense the sources in the FoV of radar), *RadioMic* can constrain sound sensing to certain areas. This can be particularly useful in a variety of settings. First, when there are multiple smart assistants nearby, they can all be triggered simultaneously, as they can all sense the same sound. On the other hand, for improved user experience, each smart assistant can be triggered for a source in a particular region. Second, private areas within a room can be built. For example, one may prefer the voice assistants to be disabled when they are doing privacy-sensitive work on their desktops. Instead of relying on a timing-based schedule, *RadioMic* can rely on a location-based approach, and can monitor the room partially for activating the smart assistant. We believe there are many interesting directions along these lines, and they can be explored fully by the researchers, with the help of the findings in this thesis.

6.2.2 *RadioSES*: mmWave-Based Audioradio Sound Enhancement and Separation System System

Our system in Chapter 4 relies on a multimodal (*audioradio*) system to solve inherent problems in the audio domain, and shows great potential in including the radio modality for speech enhancement and separation tasks. Again, even though *RadioSES* is a practical system in terms of computational efficiency, real-time properties (*e.g.* processing delay), and practical assumptions, there are a variety of directions that can be pursued, some of which are listed below.

1. **Motion Limitations:** In general, mmWave-based methods include the assumption of static bodies, as separating the motion signature from the vital sign's signature is a difficult task. On the other hand, to improve the performance in speech enhancement or separation, an algorithm does not need to capture the vocal folds clearly during the entire time. Therefore, an improved algorithm can focus on (or *attend to*) extracting vocal folds vibration when there is minimal motion, and use that prior information later on to improve the system. Such a system can be more robust against different types of motions and can make the system more practical.

2. **Multimodal Systems:** *RadioSES* is a system inspired by the audiovisual systems for speech enhancement and separation. We propose to use an audioradio system to solve the problems of audio-only systems, and we assert that an mmWave-based system has benefits over an audiovisual system, especially if there are constraints on computational complexity or privacy concerns. On the other hand, some applications do not have those limitations, and we believe that they can benefit from incorporating multiple modalities.

Namely, combining both a video and a radio stream can potentially enable high-fidelity speech enhancement and separation. Furthermore, this combination may help with some of the practical limitations. As an example, the radio stream cannot capture the vibration of the vocal folds well when the motion is strong, whereas video processing algorithms are more robust to motion. On the other hand, vision-based methods fail to operate when there are occlusions (*e.g.* face mask), where the radar devices are more robust. Combining multiple modalities can relax the practicality constraints even further and enable more interesting applications.

6.2.3 *RadioVAD*: mmWave Based Robust Voice-Activity Detection System

Our system in Chapter 5 explores radio-based or audioradio voice activity detection for improving robustness in practical systems. Compared to the *RadioMic* and *RadioSES*, *RadioVAD* is the most practical system with the least amount of constraints. On the other hand, there is still room for improvement to make the system more applicable in real-world scenarios. These include the following:

1. **Improved Responsiveness:** As discussed in Chapter 5, *RadioVAD* can detect voice activity detection with minimal delay in many cases. Although the delay is comparable to the other baselines, due to the fact that VAD is the first step for many applications, the delay requirements are usually much stricter than other applications (*e.g.* humans get annoyed by signal delays of more than 30ms). Therefore, there is room for improvement to make the system more realistic, in terms of reducing detection delay.

2. **Additional Applications:** As mentioned before, a voice activity detection system is usually the first step to many high-order tasks, and it can be incorporated in many systems, such as in *RadioSES* to relax the constraint of highly-overlapping speech, or in *RadioMic* to further improve the voice-based vibration detection in challenging (*e.g.* with motion) environments. *RadioVAD* can also be fine-tuned for hotword detection, or user identification, and trigger based on a particular passphrase or the identity of the user.

6.2.4 Overall Future Work and Concluding Remarks

With the widespread availability of acoustic/microphone-based IoT devices in recent years, the limitations of microphones have been more eminent. On the other hand, wireless has been another medium for sensing over the last few years, starting from using existing communication systems such as WiFi, to more customized by also widely adopted devices in UWB and mmWave bands. This emerging modality has a great potential for many interesting applications, which have only been explored partially. We believe, wireless-based sensing can be as ubiquitous as video in the future, and we have introduced in this thesis different systems to advance those systems further. We have addressed a multitude of limitations of current microphone-based sensing systems by incorporating wireless sensing into these smart devices. We expect to see more research work towards this goal in the future, and we believe the following are some of the future research directions to enable advanced **audioradio** systems.

1. In this thesis, all of the proposed systems rely on mmWave-based wireless sensing. Many of these applications can also be targeted by 60 GHz WiFi (802.11ad) and with the recent wireless sensing standard under development (802.11bf). Furthermore, similar results can

be potentially achieved by UWB sensing devices (*e.g.* UWB radars), where a natural extension would be using UWB *communication* devices. Other custom-built wireless devices can achieve similar goals, assuming they have enough resolution. In short, the systems in this thesis can be extended to other *wireless* medium, which is expected to be absorbed more in the next couple of years. The commonality in this medium is using *radio-frequency* waves, and this is the origin of our common title *radio* of all our system.

2. As explained in Chapter 1, apart from sound and vibration sensing, there have been many interesting applications using mmWave radar devices for human and environment monitoring. Even though there is a great promise to use mmWave for these applications, most of the ideas have been in the proof-of-concept stage, and there has been a limited absorption of mmWave sensing in the industry. By demonstrating the new capabilities of mmWave-based sensing in this thesis, we believe we provided more justification to include these sensors in smart devices. In order to push mmWave research further, there is a need from the industry to support and implement these systems in products, and that would enable further applications. One of the major limitations of mmWave-based sensing is lack of large datasets (compared to audiovisual systems), and this can only be mitigated by extensive deployment and data collection. Large datasets and availability would carry the mmWave research further, and we are excited about the numerous capabilities of mmWave sensing.

Bibliography

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 33, no. 4, pp. 79:1–79:10, 2014.

[2] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "UWHear: Through-wall extraction and separation of audio vibrations using wireless signals," in *Proc. of the ACM SenSys*, p. 1–14, 2020.

[3] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmVib: Micrometer-level vibration measurement with mmwave radar," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[4] F. Zhang, C. Wu, B. Wang, H.-Q. Lai, Y. Han, and K. J. R. Liu, "Widetect: Robust motion detection with a statistical electromagnetic model," *ACM IMWUT*, vol. 3, Sept. 2019.

[5] X. Liu, J. Cao, S. Tang, and J. Wen, "Wi-sleep: Contactless sleep monitoring via wifi signals," in *2014 IEEE Real-Time Systems Symposium*, pp. 346–355, IEEE, 2014.

[6] F. Zhang, C. Wu, B. Wang, M. Wu, D. Bugos, H. Zhang, and K. J. R. Liu, "Smars: Sleep monitoring via ambient radio signals," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 217–231, 2019.

[7] F. Zhang, C. Chen, B. Wang, and K. J. R. Liu, "Wispeed: A statistical electromagnetic approach for device-free indoor speed estimation," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2163–2177, 2018.

[8] Y. Hu, M. Z. Ozturk, F. Zhang, B. Wang, and K. J. R. Liu, "Robust device-free proximity detection using wifi," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7918–7922, IEEE, 2021.

[9] F. Zhang, C. Chen, B. Wang, H.-Q. Lai, Y. Han, and K. J. R. Liu, "Wiball: A time-reversal focusing ball method for decimeter-accuracy indoor tracking," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 4031–4041, 2018.

[10] "Texas instruments, iwr1443," 2020.

[11] "Decawave dw1000," 2020.

[12] "Netgear, nighthawk x10 smart wifi router," 2021.

[13] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges," *Wireless networks*, vol. 21, no. 8, pp. 2657–2676, 2015.

[14] C. Wu, F. Zhang, B. Wang, and K. J. R. Liu, "mSense: Towards mobile material sensing with a single millimeter-wave radio," in *ACM IMWUT*, Sep 2020.

[15] R. Thomä, T. Dallmann, S. Jovanoska, P. Knott, and A. Schmeink, "Joint communication and radar sensing: An overview," in *2021 15th European Conference on Antennas and Propagation (EuCAP)*, pp. 1–5, 2021.

[16] "Soli radar-based perception and intercation in pixel 4," 2020.

[17] "Contactless sleep sensing in nest hub with soli," 2021.

[18] I. V. Mikhelson, S. Bakhtiari, T. W. Elmer II, and A. V. Sahakian, "Remote sensing of heart rate and patterns of respiration on a stationary subject using 94-ghz millimeter-wave interferometry," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 6, pp. 1671–1677, 2011.

[19] F. Wang, F. Zhang, C. Wu, B. Wang, and K. J. R. Liu, "Vimo: Vital sign monitoring using commodity millimeter wave radio," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[20] F. Wang, X. Zeng, C. Wu, B. Wang, and K. J. R. Liu, "mmHRV: Contactless heart rate variability monitoring using millimeter-wave radio," *IEEE Internet of Things Journal*, vol. 8, no. 22, pp. 16623–16636, 2021.

[21] U. Ha, S. Assana, and F. Adib, "Contactless seismocardiography via deep learning radars," in *Proceedings of the 26th ACM Mobicom*, (New York, NY, USA), Association for Computing Machinery, 2020.

[22] X. Yang, J. Liu, Y. Chen, X. Guo, and Y. Xie, "MU-ID: Multi-user identification through gaits using millimeter wave radios," in *Proc. of the IEEE INFOCOM 2020*, pp. 2589–2598, 2020.

[23] Z. Meng, S. Fu, J. Yan, H. Liang, A. Zhou, S. Zhu, H. Ma, J. Liu, and N. Yang, "Gait recognition for co-existing multiple people using millimeter wave sensing," in *Proceedings of AAAI*, vol. 34, pp. 849–856, 2020.

[24] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "GaitCube: Deep data cube learning for human recognition with millimeter-wave radio," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[25] E. Hayashi, J. Lien, N. Gillian, L. Giusti, D. Weber, J. Yamanaka, L. Bedal, and I. Poupyrev, "Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[26] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proc. of the ACM SIGCOMM*, pp. 269–282, 2015.

[27] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st ACM Annual International Conference on Mobile Computing and Networking*, pp. 65–76, 2015.

[28] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3d human pose construction using wifi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pp. 1–14, 2020.

[29] B. Xie and J. Xiong, "Combating interference for long range lora sensing," in *Proc. of the ACM SenSys 2020*, pp. 69–81, 2020.

[30] L. Chen, J. Xiong, X. Chen, S. I. Lee, K. Chen, D. Han, D. Fang, Z. Tang, and Z. Wang, "Widesee: Towards wide-area contactless wireless sensing," in *Proc. of the 17th ACM SenSys*, pp. 258–270, 2019.

[31] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proc. of the ACM MobiSys 2019*, pp. 313–325, 2019.

[32] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. J. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE SPM*, vol. 35, no. 3, pp. 59–80, 2018.

[33] K. J. R. Liu and B. Wang, *Wireless AI: Wireless Sensing, Positioning, IoT, and Communications*. Cambridge University Press, 2019.

[34] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[35] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proc. of the 33rd ACM CHI*, pp. 837–846, 2015.

[36] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *Proceedings of the 23rd ACM Annual International Conference on Mobile Computing and Networking*, pp. 211–220, 2016.

[37] F. Wang, F. Zhang, C. Wu, B. Wang, and K. J. R. Liu, "Vimo: Multiperson vital sign monitoring using commodity millimeter-wave radio," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1294–1307, 2021.

[38] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *Proceedings of IEEE INFOCOM*, pp. 1472–1480, IEEE, 2015.

[39] J. Xiao, K. Wu, Y. Yi, L. Wang, and L. M. Ni, "Pilot: Passive device-free indoor localization using channel state information," in *Proc. of the IEEE ICDCS 2013*, pp. 236–245, 2013.

[40] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals," in *Proceedings of the ACM UbiComp 2016*, pp. 363–373, 2016.

[41] Z.-W. Li, "Millimeter wave radar for detecting the speech signal applications," *International Journal of Infrared and Millimeter Waves*, vol. 17, no. 12, pp. 2175–2183, 1996.

[42] Y. Rong, S. Srinivas, A. Venkataramani, and D. W. Bliss, "Uwb radar vibrometry: An rf microphone," in *Proc. of the IEEE ACSSC*, pp. 1066–1070, 2019.

[43] E. Guerrero, J. Brugués, J. Verdú, and P. de Paco, "Microwave microphone using a general purpose 24-ghz fmcw radar," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1–4, 2020.

[44] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "Sound recovery from radio signals," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[45] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "RadioMic: Sound sensing via mmwave signals," *CoRR*, vol. abs/2108.03164, 2021.

[46] F. Chen, S. Li, Y. Zhang, and J. Wang, "Detection of the vibration signal from human vocal folds using a 94-ghz millimeter-wave radar," *MDPI Sensors*, p. 543, Mar 2017.

[47] S. Li, Y. Tian, G. Lu, Y. Zhang, H. Lv, X. Yu, H. Xue, H. Zhang, J. Wang, and X. Jing, "A 94-ghz millimeter-wave sensor for speech signal acquisition," *Sensors*, vol. 13, no. 11, pp. 14248–14260, 2013.

[48] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proc. of the ACM MobiSys 2019*, pp. 14–26, 2019.

[49] Y. Dong and Y.-D. Yao, "Secure mmwave-radar-based speaker verification for iot smart home," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3500–3511, 2020.

[50] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren, *et al.*, "VocalPrint: exploring a resilient and secure voice authentication via mmwave biometric interrogation," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pp. 312–325, 2020.

[51] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals," in *Proc. of the ACM SenSys*, (New York, NY, USA), p. 97–110, Association for Computing Machinery, 2021.

[52] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "Toward mmwave-based sound enhancement and separation," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6852–6856, 2022.

[53] M. Z. Ozturk, C. Wu, B. Wang, M. Wu, and K. J. R. Liu, "RadioSES: mmwave-based audioradio speech enhancement and separation system," 2022.

[54] A. Izzo, L. Ausiello, C. Clemente, and J. J. Soraghan, "Loudspeaker analysis: A radar based approach," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1223–1237, 2020.

[55] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proceedings of the 21st ACM Annual International Conference on Mobile Computing and Networking*, p. 130–141, 2015.

[56] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. Ni, "We can hear you with wi-fi!," 2014.

[57] A. M. Eid and J. W. Wallace, "Ultrawideband speech sensing," *IEEE Antennas and Wireless Propagation Letters*, vol. 8, pp. 1414–1417, 2009.

[58] P. Birkholz, S. Stone, K. Wolf, and D. Plettemeier, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2404–2411, 2018.

[59] D. L. Jennings and D. W. Ruck, "Enhancing automatic speech recognition with an ultrasonic lip motion detector," in *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 868–871 vol.1, 1995.

[60] K. Kalgaonkar and B. Raj, "Ultrasonic doppler sensor for speaker recognition," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4865–4868, 2008.

[61] A. R. Toth, K. Kalgaonkar, B. Raj, and T. Ezzat, "Synthesizing speech from doppler signals," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4638–4641, 2010.

[62] K.-S. Lee, "Speech enhancement using ultrasonic doppler sonar," *Speech Communication*, vol. 110, pp. 21–32, 2019.

[63] K. Sun and X. Zhang, "UltraSE: Single-channel speech enhancement using ultrasound," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, (New York, NY, USA), p. 160–173, 2021.

[64] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Real-time passive sound recovery from light bulb vibrations." Cryptology ePrint Archive, Report 2020/708, 2020.

[65] R. P. Muscatell, "Laser microphone," Oct. 23 1984. US Patent 4,479,265.

[66] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: Eavesdropping via lidar sensors," in *Proc. of the ACM SenSys 2020*, pp. 354–367, 2020.

[67] G. Galatas, G. Potamianos, and F. Makedon, "Audio-visual speech recognition incorporating facial depth information captured by the kinect," in *Proc. of the EURASIP EUSIPCO*, pp. 2714–2717, 2012.

[68] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proc. of the ACM MobiSys 2015*, pp. 301–315, 2015.

[69] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *USENIX Security*, (San Diego, CA), pp. 1053–1067, Aug. 2014.

[70] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proc. of the ACM MobiSys 2016*, p. 57–69, 2016.

[71] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[72] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[73] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization.," in *Proc. of the Interspeech 2006*, vol. 2, pp. 2–5, Citeseer, 2006.

[74] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space.," in *Proc. of the Interspeech 2006*, 2006.

[75] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[76] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 79–82, 2001.

[77] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[78] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[79] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[80] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. of the AAAI Conference on Artificial Intelligence*, no. 05, pp. 9458–9465, 2020.

[81] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[82] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. of the Interspeech 2017*, 2017.

[83] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[84] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path Rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, IEEE, 2020.

[85] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[86] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2021.

[87] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[88] M. Kolbaek, D. Yu, Z. Tan, and J. H. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1901–1913, 2017.

[89] R. Lu, Z. Duan, and C. Zhang, "Audio–visual deep clustering for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[90] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[91] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 246–250, IEEE, 2017.

[92] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. of the ICML 2020*, 2020.

[93] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM TOG*, vol. 37, July 2018.

[94] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Proc. of the Interspeech 2018*, pp. 3244–3248, 2018.

[95] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2019*, pp. 667–673, IEEE, 2019.

[96] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7539–7548, 2019.

[97] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," in *Proc. of the Interspeech 2020*, pp. 3481–3485, 10 2020.

[98] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[99] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[100] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[101] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[102] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *IEEE ICASSP 2013*, pp. 7378–7382, 2013.

[103] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-end domain-adversarial voice activity detection," *arXiv preprint arXiv:1910.10655*, 2019.

[104] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal vad: Speaker-conditioned voice activity detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pp. 433–439, 2020.

[105] S. Ding, R. Rikhye, Q. Liang, Y. He, Q. Wang, A. Narayanan, T. O'Malley, and I. McGraw, "Personal vad 2.0: Optimizing personal voice activity detection for on-device speech recognition," *arXiv preprint arXiv:2204.03793*, 2022.

[106] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[107] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[108] P. Liu and Z. Wang, "Voice activity detection using visual information," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–609, IEEE, 2004.

[109] S. P. Arjunan, H. Weghorn, D. K. Kumar, and W. C. Yau, "Vowel recognition of english and german language using facial movement (semg) for speech control based hci," in *Proceedings of the HCSNet Workshop on Use of Vision in HCI-Volume 56*, pp. 13–18, 2006.

[110] F. Faubel, M. Georges, K. Kumatani, A. Bruhn, and D. Klakow, "Improving hands-free speech recognition in a car through audio-visual voice activity detection," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 70–75, 2011.

[111] S. Guy, S. Lathuilière, P. Mesejo, and R. Horaud, "Learning visual voice activity detection with an automatically annotated dataset," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4851–4856, IEEE, 2021.

[112] M. Shahid, C. Beyan, and V. Murino, "S-VVAD: Visual voice activity detection by motion segmentation," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2331–2340, 2021.

[113] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, 2019.

[114] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *IEEE ICASSP 2020*, pp. 4492–4496, IEEE, 2020.

[115] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media/multi-modal resources with elan.," in *LREC*, pp. 2065–2068, 2004.

[116] R. Hu and B. Raj, "A robust voice activity detector using an acoustic doppler radar," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pp. 319–324, IEEE, 2005.

[117] K. Kalgaonkar, R. Hu, and B. Raj, "Ultrasonic doppler sensor for voice activity detection," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, 2007.

[118] I. V. McLoughlin, "Super-audible voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1424–1433, 2014.

[119] A. Honarmandi Shandiz and L. Tóth, "Voice activity detection for ultrasound-based silent speech interfaces using convolutional neural networks," in *International Conference on Text, Speech, and Dialogue*, pp. 499–510, Springer, 2021.

[120] L. Ding, M. Ali, S. Patole, and A. Dabak, "Vibration parameter estimation using fmcw radar," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2224–2228, 2016.

[121] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F (Radar and Signal Processing)*, no. 5, pp. 343–350, IET, 1992.

[122] G. Fant, *Acoustic theory of speech production*. No. 2, Walter de Gruyter, 1970.

[123] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proc. of the ACM MobiSys 2017*, pp. 2–14, 2017.

[124] F. J. Fahy, *Foundations of engineering acoustics*. Elsevier, 2000.

[125] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[126] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4214–4217, 2010.

[127] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Ellis Horwood Series in Artificial Intelligence, Prentice Hall, 1988.

[128] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proc. of the ACM CCS 2017*, pp. 103–117, 2017.

[129] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE APSIPA ASC*, pp. 1–5, 2014.

[130] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proc. of the ACM SIGCOMM 2018*, pp. 267–281, 2018.

[131] F. Zhang, C. Wu, B. Wang, and K. J. R. Liu, "mmEye: Super-resolution millimeter wave imaging," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6995–7008, 2021.

[132] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[133] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4395–4399, 2015.

[134] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2018.

[135] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[136] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[137] "Kron tech. chronos 1.4 high speed camera," 2021.

[138] "2020: The year of the affordable ultra high-speed cameras," 2021.

[139] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 175–179, IEEE, 2019.

[140] A. Chan, L. Mongeau, and K. Kost, "Vocal fold vibration measurements using laser doppler vibrometry," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1667–1676, 2013.

[141] G. Ramachandran and M. Singh, "Three-dimensional reconstruction of cardiac displacement patterns on the chest wall during the p, qrs and t-segments of the ecg by laser speckle inteferometry," *Medical and Biological Engineering and Computing*, vol. 27, no. 5, pp. 525–530, 1989.

[142] R. Khanna, D. Oh, and Y. Kim, "Through-wall remote human voice recognition using doppler radar with transfer learning," *IEEE Sensors Journal*, vol. 19, no. 12, pp. 4571–4576, 2019.

[143] K. V. Mishra, M. Bhavani Shankar, V. Koivunen, B. Ottersten, and S. A. Vorobyov, "Toward millimeter-wave joint radar communications: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 100–114, 2019.

[144] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[145] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. of the Interspeech 2019*, Sept. 2019.

[146] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[147] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, no. 7-8, pp. 667–677, 2007.

[148] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1066–1078, 2016.

[149] "The next amazon echo could use radar to track your sleep," 2021.

[150] T. Higuchi, T. Yoshioka, and T. Nakatani, "Optimization of speech enhancement front-end with speech recognition-level criterion.," in *Interspeech*, pp. 3808–3812, 2016.

[151] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[152] M. A. Richards, *Fundamentals of Radar Signal Processing*. Professional Engineering, McGraw-Hill Education, 2005.

[153] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via rf body reflections," in *Proc. of the USENIX NSDI 2015*, pp. 279–292, 2015.

[154] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, p. 226–231, AAAI Press, 1996.

[155] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[156] T. Zheng, Z. Chen, S. Zhang, C. Cai, and J. Luo, "More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar," in *Proc. of the ACM SenSys 2021*, (New York, NY, USA), p. 111–124, Association for Computing Machinery, 2021.

[157] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[158] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, vol. 4, pp. 2861–2866 vol.4, 2001.

[159] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, IEEE, 2019.

[160] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[161] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020.

[162] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. of the Interspeech 2018*, 2018.

[163] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[164] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. of the ICML 2011*, 2011.

[165] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[166] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *Proc. of the IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2020.

[167] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498–1507, 2009.

[168] M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: an image classification approach," *Sensors*, vol. 20, no. 1, p. 172, 2020.

[169] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Proc. Interspeech 2019*, pp. 4295–4299, 2019.

[170] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of lombard effect," *Speech Communication*, vol. 115, pp. 38–50, 2019.

[171] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[172] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.

[173] X. Xiang, X. Zhang, and H. Chen, "Acquisition and enhancement of remote human vocal signals based on doppler radar," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20348–20361, 2021.

[174] K. Ahuja, A. Kong, M. Goel, and C. Harrison, *Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems*, p. 1121–1131. New York, NY, USA: Association for Computing Machinery, 2020.

[175] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier." `https://github.com/snakers4/silero-vad`, 2021.

[176] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, pp. 1015–1021, Springer, 2006.