RADIOSES: mmWave-Based Audioradio Speech Enhancement and Separation System

Muhammed Zahid Ozturk, Student Member, IEEE, Chenshu Wu, Senior Member, IEEE, Beibei Wang, Senior Member, IEEE, Min Wu, Fellow, IEEE, and K. J. Ray Liu, Fellow, IEEE

Abstract—Speech enhancement and separation have been a long-standing problem, especially with the recent advances using a single microphone. Although microphones perform well in constrained settings, their performance for speech separation decreases in noisy conditions. In this work, 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. Similar to the audiovisual methods, RADIOSES provides significant performance improvements (e.g. 3 dB gains in SISDR, when compared with the corresponding audio-only method), along with the benefits of lower computational complexity and better privacy preservation.

Index Terms—Wireless sensing, mmWave sensing, speech enhancement, speech separation, deep learning.

I. 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 [1]. SES capability for computers is in 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 recent years with the help of deep learning, especially when there is not much background noise [2]. 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. Although these problems can be solved for clean mixtures, by clustering-based methods [3] and permutation invariant training (PIT) [4], their performance can decrease with noisy mixtures. Overall, audio-only approaches suffer from these ill-posed problems inherently.

To overcome the problems and enhance SES, multimodal systems have been introduced to exploit readily available information beyond audio, such as video [5], [6]. Similar to human perception, which also uses lip motion and facial information [7], 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 [8], or other information, such as voice activity detection [9], or pitch [10] 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 the 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 [11], [12], Amazon Alexa [13], 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 [14], measuring machinery and object vibration [15], [16], or extracting vocal folds vibration [17]. In particular, it has been used to estimate pitch and detect voice activity [17], reconstruct speech to some extent [18], [19], as well as enhance speech recognition for a single speaker [20]. Yet no existing work has explored utilizing both modalities for joint SES tasks.

2329-9290 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
With this motivation, we develop an audioradio\(^1\) speech enhancement and separation system to solve the aforementioned problems and improve overall performance. Building an audioradio SES system faces multiple challenges. First, in order to solve the number of sources estimation problem, robust and efficient source detection, and tracking methods are 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 by motion, with motion usually being the stronger effect. Third, different from the rich literature on 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 Speech Enhancement and Separation system. As illustrated in Fig. 1, RADIOSES can detect, localize, and estimate the 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 RADIOSES\(_1\)ET, 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 the robustness and generalizability of RADIOSES\(_1\)ET 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 [21]), RADIOSES brings around 3 dB improvements for separating noisy mixtures, along with the benefits of correctly 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 [22] 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 propose to use learnable encoders for the time-frequency representation of radio signals, similar to the learnable encoders in the speech domain, 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 paper follows a literature review in Section II, and a preliminary in Section III. Section IV presents an overview, with detailed design in Section V and Section VI. We give dataset and implementation details in Section VII-B, and present the results in Section VIII. Last, we discuss in Section IX and conclude in Section X.

II. RELATED WORK

**Audio-only Methods**: Traditional methods, such as computational auditory scene analysis (CASA) [23] with pitch-estimation based separation [24], nonnegative matrix factorization (NMF) [25], or probabilistic methods [26] cannot generalize well to unseen speakers [27], which is a major limiting factor for their performance.

Deep learning based methods outperformed classical approaches recently [27]. Instead of estimating output representation directly, these methods usually estimate a mask that is multiplied with the input. Some masks used as training targets are binary mask [28], STFT spectral mask [29], and complex ratio mask [30]. PHASEN [31] estimates amplitude and phase masks separately. In SEGAN [32], a time-domain SE system

---

\(^1\)We combine audio and radio words as audioradio to refer to a multimodal system consisting of both modalities, similar to the word audiovisual.
using generative adversarial networks has been proposed. ConvTasNet [33] performs better than the ideal ratio mask for SS, with an adaptive/learnable encoder, instead of classical STFT. Later on, the fully convolutional layers in [33] are replaced by dual-path RNN (DPRNN-TasNet) [21], dual-path transformer network (DPTNET) [34], and fully attention layers in SepFormer [35].

Source association and tracking problems can be solved with frame-level PIT [4] and utterance-level PIT [36]. 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 [37]. The number of sources can be estimated by deep clustering [3] or deep attractor networks [38]. However, these models still have the source tracking problem over a long time, which is started to be addressed recently [39].

Multimodal Methods: Vision-based works use different features as input, such as face embeddings [5], lip embeddings, [6] or optical flow [37]. These methods use STFT representation, although time-domain processing is also possible [40]. [41] estimates faces from the speech signals, whereas [8] uses pictures of speakers 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 [42], speaker recognition [43], and speech enhancement [44]. UltraSE [45] uses deep learning to enhance single-speaker speech. Ultrasound signals can only work at a short range (e.g. 15 cm) and are too coarse to measure fine-grained vocal folds vibration [45].

Last, laser-Doppler vibrometers (LDV) can also be used for speech enhancement from a distance [46], speech recognition [47], and speech reconstruction [48]. LDVs are usually very expensive, heavy, and power-hungry, and therefore not very practical.

Wireless Sensing: Recently, wireless sensing has become an emerging phenomenon [49], [50], with multiple applications to gait monitoring [51], [52], vibration monitoring [16], and vital signs monitoring [14], [53], [54]. mmWave devices have enabled monitoring of \( \mu m \) level displacement on object surfaces and can capture vocal folds vibration remotely [17]. Furthermore, they can recover sound from human throats [18], speaker diaphragms [55], or passive object surfaces, such as a piece of aluminum [19]. Recently, mmWave information from the speaker has also been used for speaker verification [56], speaker recognition, [57] or voice activity detection [58]. Although these works can reject interfering sound by sensing its source, they capture limited and low-quality sound. Such systems do not employ one of the most common sensors available, microphones, to further improve the quality. A recent work, WaVoice [20], fuses both modalities for speech recognition, but can only recognize commands of a single user, and is not suitable for speech separation. To that end, RADIOSES explores a multimodal speech enhancement and separation system using radio signals. Furthermore, to the best of our knowledge, RADIOSES is the first work to utilize adaptive encoders for the time-frequency representation of radio signals, instead of spectrograms, which is the typical method for all radio-based sensing systems.

III. 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 \( \mu m \) level vibration displacement on the surface of the human body, especially in the throat region, and this displacement changes the amplitude and phase of the returned complex-valued radar signals. As shown in Fig. 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-quality speech enhancement and separation. We note that, although Fig. 2 includes spectrograms for illustration, RADIOSES uses learnable encoders for time-frequency representation of both audio and radio modalities.

IV. SYSTEM OVERVIEW

As an overview, RADIOSES requires a device with mmWave sensing capabilities, and a microphone, (e.g. [11], [12]). 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 the radar and microphone to be colocated, as in [11]. We expect the speaking subjects to be in front of the radar and to not block each other completely. In addition, although radars can sense in non-line-of-sight (NLOS) conditions, we only investigate line-of-sight (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 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 a clean speech or passed through a speech-to-text engine to convert into commands.

The first main block of RADIOSES, source detection and localization in Fig. 3, is explained briefly in Section V, whereas the second block, the deep learning module is further detailed in Section VI.

V. RADIO FEATURE EXTRACTION

As shown in Fig. 3, the goal of the radio feature extraction module is to output individual radio streams from sources in the environment. To achieve that, we adopt a variety of methods in an efficient pipeline to detect and locate targets. Unlike existing works, such as [19], [55], RADIOSES does not rely on a spectrogram-based metric to locate 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 [59]. As mmWave devices usually have multiple antennas, we define the CIR at the m-th antenna $h_m(t)$ as:

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

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 $\tau_r$ denote the complex amplitude and the propagation delay of the r-th range bin, and $\epsilon$ denotes the additive noise, respectively. Here, the range resolution $\Delta R$ can be inferred from the time resolution, $\Delta t$, which is inversely proportional to bandwidth (corresponding to 4.26 cm for our device). Therefore, a separate stream from very close targets can be extracted. The CIR in (1) is captured repeatedly during sensing and is time dependent. To simplify (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 [60]. Range-azimuth CIR is denoted by $h_{r,\theta}(t)$, where $\theta$ represents the azimuth angle. Since our virtual antenna array elements are placed $d = \lambda/2$ apart, where $\lambda$ is the wavelength, $h_{r,\theta}(t)$ can be given as:

$$h_{r,\theta}(t) = s^H(\theta)h_{m,r}(t) + \epsilon(t), \quad (2)$$

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

$$s_m(\theta) = \exp\left(-j2\pi \frac{d \sin \theta}{\lambda}\right), \quad (3)$$

and the channel vector is $h_{m,r}(t) = [h_{1,r}(t), h_{2,r}(t), \ldots, h_{M,r}(t)]$, 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 (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, a constant false alarm rate (CFAR) detector [61], which adaptively estimates the background noise for different bins and thresholds for each range-azimuth bin accordingly. As shown in Fig. 4(a), the 2D CFAR window is denoted with $C$, and the CFAR threshold is denoted with $\gamma$. This window is applied to the magnitude of the range-azimuth plane, and the corresponding range-azimuth plane is shown in Fig. 4(b). Therefore, the CFAR detection rule on the range-azimuth plane is given as:

$$B_{r,\theta}^{\text{CFAR}}(t) = \mathbb{I}\{(C * |h_{r,\theta}|)(t) > \gamma(|h_{r,\theta}(t)|)\}, \quad (4)$$

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

Clutter Removal: 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 [62], [63]. 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(c). Therefore, human detector output is $B_{r,\theta}^{\text{stat}} = \mathbb{I}\{V_{r,\theta}(t) < \eta_{\text{stat}}\}$. Furthermore, bodies with excessive motion can also be filtered using a similar approach, and we reject those by: $B_{r,\theta}^{\text{mov}} = \mathbb{I}\{V_{r,\theta}(t) > \eta_{\text{mov}}\}$, where $\eta_{\text{stat}}(r, \theta) \triangleq \frac{\eta_{\text{stat}}(\theta)}{\frac{\pi^2 \cos^2(\theta)}{2304}}$, $H_{\text{mov}}(r, \theta) \triangleq \frac{\eta_{\text{mov}}(\theta)}{\frac{\pi^2 \cos^2(\theta)}{2304}}$. $\eta_{\text{stat}}$ and $\eta_{\text{mov}}$ are empirically found thresholds. The minimum and maximum variances are defined with respect to $(r, \theta)$, 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 the intersection of all binary maps, i.e. $B_{r,\theta}(t) = \{ B_{\text{CFAR}} \cap B_{\text{stat}} \cap B_{\text{mov}} \} (t)$, as shown in Fig. 4(d).

**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 [64]. The parameters for DBSCAN are set empirically, and an example of clustering is shown in Fig. 4(e). 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 [65] 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 [66]. 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. The output dimensionality of the radar signals is $2 \times 1000$ at 16bits for a 1-second stream, which is lower than the microphone and typical video streams.

### VI. AUDIO/RADIO 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 VI-A and then detail our design in Section VI-B.

#### A. Background and Design Rationale

**Background**: Usually, an SES model follows the architecture in Fig. 5, with an encoder, masker, and a decoder block [67]. 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 [68]. The performance can be increased by using more optimal masks (such as complex ratio mask [30]); 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 [33].

**Design Rationale**: RADIOSES uses the same structure as in Fig. 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 [16], [66]. If a spectrogram representation is used as an input, not only it may not be optimal for the neural network 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 [18], [20]. Using either the real or imaginary part of the signal (as in [55]) or combining both parts optimally with a linear projection [19] also loses important signal content. Based on this, RADIOSES uses an adaptive front-end for radio streams.

To make RADIOSES work with raw radio inputs, we apply random rotation in the in-phase/quadrature (IQ) plane, as proposed by previous work [66]. However, unlike [66], 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 VIII. We select the cutoff frequency of the high pass filter at 90 Hz in order not to filter pitch frequency. Afterward, the radio signals are encoded with an adaptive encoder, as explained in Section VI-B.

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 main issue with fully convolutional models [21]. 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.

#### B. RADIOSESNET Design

1) **Encoders**: The audio encoder of RADIOSESNET consists of a 1D convolutional layer, with kernel size 16, and the number of kernels 256, followed by ReLU nonlinearity and layer

2We refer the reader to [16] for IQ representation of the returned signals, and to [66] for discussion and introducing random rotation.
normalization. Radio channel uses another 1D convolutional layer, nonlinearity, and normalization, with 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. The output of the audio and radio encoders are represented with $X_i \in \mathbb{R}^{N_i \times L_i}$, with $*$ in $(a, r)$, for audio and radio stream, where we drop the index $i$ for simplicity. Here, $N_i$ represents the number of features, and $L_i$ represents the number of time samples of encoded representation.

2) Masker: Both encoded modalities are combined to estimate the masks for each source, as illustrated in the masker of Fig. 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 [21], where an example of DPRNN workflow is presented in Fig. 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 $\tilde{X}_i \in \mathbb{R}^{N_i \times K_i \times S_i}$, with $K_i$ and $S_i$ denoting the block length and number of blocks. The input, output representations $X_i$, $\tilde{X}_i$ and dimensionalities $N_i$, $L_i$, $K_i$, and $S_i$ are defined similarly for radio channel and given in Table I, whereas the flow for a single DPRNN process is given in Fig. 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 1, and fully connected layers are used to match the input size to the output size, which enables to repeat multiple DPRNN blocks without any size mismatches.

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 audio signals in time-domain. DPRNN blocks are converted back to a representation similar to the one at the input, by overlap-add method [21]. 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) Training: In order to train RADIOESNET, we use scale-invariant signal-to-distortion (SiSDR) [69]) as the loss function between the time-domain signals, which is given by:

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

where $a$ and $\hat{a}$ denote the target and the estimated sound signals. The 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 $10^{-6}$. Since a separate model for different numbers of users has been trained, RADIOESN switches to the appropriate model by estimating the number of sources.

5) Other Design Considerations: Complexity and causality are particularly considered in our design.

Complexity: RADIOESN has a compact design, with only 2.1M parameters. Among these, the radio stream occupies 320 k parameters, which could easily be fit on a small device. Forward
pass of a 3-second input with RadioSESNet takes 4 ms on a modern GPU with batch processing, which is only 0.4 ms 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 an algorithmic delay of around 150ms. We leave the investigation of a real-time system to the future, but RadioSES is already close to real-time processing, unlike [5], [45].

**VII. Experiment and Implementation**

**A. Data Collection**

**Hardware:** We build a data collection platform, as seen in Fig. 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 with 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 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 [70]. Our speakers come from diverse backgrounds, 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 [71]). During data collection, users sit approximately 40 cm 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 [71] with the noise files from WHAM dataset [2]. We randomly select 13 users for training and 4 users (2 male, 2 female) for evaluation. The validation set includes mixtures from the remaining two users, and the unused speech of the users in the training set. After downsampling all audio files to 8 kHz, 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 [71]). 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 VIII. 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 the future.

**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 [72].
- **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 [71], 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.

**B. 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 [73] to follow standard training and evaluation protocols in monaural SES, and to borrow implementations of
existing methods, such as ConvTasNet [33] or DPRNNNet [21]. We train RADIOSESNet and DPRNNNet for 60 epochs, using a starting learning rate of $10^{-3}$, 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 [21]. 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 [66], 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 [74] to reduce mode failure.

**VIII. Evaluation**

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 and partial inputs, and conduct an ablation study.

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

- **SiSDR** [69]: 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** [75]: 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** [76]: Short time intelligibility metric, correlates with the word error rate, reported from 0 to 1.

- **PESQ** [77]: 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 [18]. 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 estimates an upper limit of its performance, since oracle phase is not available in real noisy speech. Another recent work [20] 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 the sustained performance of RADIOSES. We include ConvTasNet [33], one of the first adaptive-encoder based systems that outperformed STFT-based masks. Second, we include DPRNNNet, which is the audio-only baseline of RADIOSES. DPRNNNet has been shown to outperform ConvTasNet significantly and can be considered as state-of-the art. Third, we use SudoRMRF [78], which simplifies DPRNNNet by replacing the RNN blocks with downsampling and upsampling blocks and is shown to achieve similar performance. When training for speech separation, we use permutation invariant training (PIT) for the audio baselines.

Last, we cannot compare with UltraSE [45], as it uses ultrasound modality, and different speakers and noise dataset. Due to changes in datasets and a 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.

**A. Speech Enhancement**

In speech enhancement, RADIOSES brings improvements to the audio-only baseline methods, as shown in Table II. 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 [5]), and shows that RADIOSES learns to exploit the radio information. On the other hand, results from WaveVoiceNet suggest that 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 [18]), our phonetically rich diverse dataset (5762 unique words vs. 631 in [18]), and users. As the results are poor, we do not investigate WaveVoiceNet further in our experiments. The performance of RADIOSES matches that of ConvTasNet, with certain qualitative differences, such as a 1.5 s look-ahead in ConvTasNet, and higher computational complexity. We also note that our implementation uses a pre-trained 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 in the background can also be considered a speech enhancement problem, however, most speech enhancement methods require prior information to focus on the particular speech. If such prior information does not exist, the problem is not well defined and it is more reasonable to evaluate the performance in speech separation tasks. In order to have a fair comparison, we evaluate this case in the following sections, under speech separation.

### B. Speech Separation

In this section, we present the speech-separation results with RADIOSES, along with the previously mentioned baselines in Table III. 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 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 generalize better to unseen users, where the basic DPRNNTasNet suffers. RADIOSES not only improves signal metrics but also intelligibility and 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 3-person mixtures. As shown in Table IV, the improvements from RADIOSES is even greater for 3-person mixtures, as radio helps to extract individual streams from each user. Since the performance gains from RADIOSES increases with more users, we expect it to work well for 4 or more users. We do not test those cases for brevity.

### C. 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 permutation problem and estimating the number of sources. To that end, in Fig. 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. 10(a), 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. 10(b) 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.

### D. 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 in a variety of settings. First, to test the effect of

---

### Table III

| Model          | 2-person mix (clean) | 2-person mix (noisy) | 3-person mix (clean) | 3-person mix (noisy) |
|----------------|----------------------|----------------------|----------------------|----------------------|
|                | SiSDR | SIR | STOI | PESQ | SiSDR | SIR | STOI | PESQ | SiSDR | SIR | STOI | PESQ |
| ConvTasNet     | 11.3  | 18.5 | 0.87 | 2.53 | 6.1   | 16.8 | 0.77 | 1.78 | 4.5   | 13.4 | 1.0  | 0.9  |
| SudoRMRF       | 10.9  | 15.4 | 0.84 | 2.60 | 4.7   | 16.4 | 0.68 | 1.77 | 3.9   | 13.0 | 0.9  | 0.9  |
| RADIOSES       | 15.4  | 23.6 | 0.94 | 2.83 | 10.9  | 23.3 | 0.85 | 2.10 | 9.3   | 19.2 | 0.83 | 1.96 |

The best values for each metric are indicated in bold.

### Table IV

| Model          | 3-person mix (clean) | 3-person mix (noisy) | 3-person mix (clean) | 3-person mix (noisy) |
|----------------|----------------------|----------------------|----------------------|----------------------|
|                | SiSDR | SIR | STOI | PESQ | SiSDR | SIR | STOI | PESQ | SiSDR | SIR | STOI | PESQ |
| Input          | -3.2  | -2.8 | 0.60 | 1.37 | -4.2  | -2.8 | 0.55 | 1.30 | 4.9   | 15.7 | 0.74 | 1.68 |
| DPRNN          | 7.2   | 14.0 | 0.81 | 1.95 | 9.9   | 19.2 | 0.83 | 1.96 | 8.6   | 18.2 | 0.81 | 1.90 |
| RADIOSES       | 11.6  | 19.4 | 0.88 | 2.31 | 9.3   | 19.2 | 0.83 | 1.96 | 8.6   | 18.2 | 0.81 | 1.90 |

The best values for each metric are indicated in bold.

---

Fig. 9. Learning curve for audio-only (AO) and audioradio (AR) for separating 2-person mixtures.

Fig. 10(a). Comparison of output SiSDR between audio-only and audioradio systems for 2-person mixtures.

---

OZTURK et al.: RADIOSES: mmWave-BASED AUDIORADIO SPEECH ENHANCEMENT AND SEPARATION SYSTEM
TABLE V
PERFORMANCE WITH RESPECT TO MULTIPLE EXPERIMENTS OF SOURCES

| Exp Case | Distance | Orientation | Head Orientation |
|----------|----------|-------------|------------------|
|          | 50 cm    | 75 cm       | 100 cm           |
|          | 0°       | 15°         | 30°              |
|          | 45°      |              |                  |
| Metric   | AO AR    | AO AR AR    | AO AR AR AR 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        |
|          |          |             |                  |
|          | 4.8 8.3  | 4.3 8.2     | 4.2 8.1          |
|          | 10.6 15.1| 10.2 15.6   | 10.1 15.5        |
|          | 0.79 0.89| 0.78 0.89   | 0.78 0.88        |
|          | 2.00 2.33| 2.02 2.37   | 2.01 2.36        |
|          | 2.16 2.61| 2.11 2.46   | 2.10 2.43        |

Fig. 10. Comparison of RADIOSES with audio-only baseline in 2-person noisy mixtures.

distance, we collect user data at different distances (e.g. 75 cm) separately and create synthetic mixtures from each location. For example, we collect multiple realizations from different users at 75 cm, and apply a similar synthetic data generation pipeline as before. We normalize input data streams to the same loudness levels for a fair comparison, although minor differences between each setting is inevitable. Other experiments, such as angle or orientation are also conducted similarly. In order to show improvements, we present each setting’s performance alongside 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 4 dB 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.

1) Distance: First, we evaluate the effect of distance on the signal separation tasks, as illustrated in Fig. 11(a). As shown in Table V, RADIOSES can work robustly until the speakers are 1 m 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 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 microphones recently (e.g. [79], [80]). We note that the performance gains from the radio channel do not decrease much from 0.5 m and 1 m, 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.

2) Orientation: Second, we ask the users to sit 0.75 m away from the device and change their orientation to explore the sensing area, as illustrated in Fig. 11(b). We realize that RADIOSES can work until 45°, without any performance decrease, as presented in orientation columns of Table V. The gains from the audioradio system are consistent (e.g. \( \sim 4 \text{ dB in SiSDR} \)) 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.

3) Head Orientation: Third, we ask users to sit at 0.5 m, and rotate their heads from 0 degrees to 15 and 30 degrees, as shown in Fig. 11(c). For example, if a user sits in front of a laptop or monitor, they would naturally swing their head to see different content on the screen and 30 degrees of head rotation at 0.5 m 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 V, which indicates that RADIOSES is robust to changes in head orientation, even though the training procedure does not include explicit head-rotation data.

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. As shown in Fig. VI-D4, RADIOSES is not affected by head motion or occlusion. Furthermore, unlike certain visual enhancement methods which lose their advantage with occlusions (as noted in [81]), RADIOSES is robust against wearing a mask and can preserve the improvements compared to the audio-only method. We expect this behavior to generalize to other face masks, and clothing materials, since these usually affect the radar response minimally, as noted in many papers, such as [14], [20]. This is because vocal folds vibration...
are mostly extracted from the body and throat, and mmWave signals can penetrate through clothing materials easily. 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. We also do not expect hair motion to affect the performance of RadioSES, since the surface area and reflected signal energy from the hairs are much smaller and weaker than that of the human body or throat.

E. 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. 11(d) for the multiple speaker case. Although making a real-world system based on multimodal sensing, and end-to-end deep learning frameworks involves additional challenges due to Lombard effect [82], 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 [18]), 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 VI, 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/.

F. 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 13(a). At larger distances, radio signals 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 -5 dB. 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 2 s, 1.5 s, 1 s, and 0.5 s. Such configurations can be used when there are power requirements or package loss in the radio stream. As shown in Fig. 13(b), RadioSES can still help with speech separation tasks and improve the performance, compared to the audio-only baseline, when there is at least 1 s 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 1.5 s 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.

G. 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 VII, RADIOSES can still outperform the audio baseline by a large margin and improve the performance, even 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 the audio-only system is larger, and benefits of having the partial radio signals are clear.

H. 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 VIII, 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.

TABLE VII

| Case | 2-person (noisy) | 3-person (noisy) |
|------|------------------|------------------|
| Metric | AO, AR(1), AR(2) | AO, AR(1), AR(2) |
| SSDR  | 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 VIII

| Model | SSDR |
|-------|------|
| RADIOSESNET | 15.4 |
| w/o Radio DPRNN | 15.2 |
| w/o Audio DPRNN | 13.5 |
| w/o HP filter | 4.8 |

IX. DISCUSSION

In this work, 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 the vibration of other sources, such as guitars [19], or machinery [16]. 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 [83].

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 [67] 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 the audiovisual domain [84], 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 [66]. A more thorough system should support medium and high levels of source motion. To that end, the coherent combining of multiple vocal fold bins from person point clouds (e.g. [14]), or deep learning [66] can be some interesting ideas to support multiple moving targets.

Sensing Distance: Our experiments indicate that RADIOSES can work robustly until the speakers are 1 m 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. 2.5 m) for people with low pitch (e.g. males). To support all users, we limited the practical range to 1 m, much larger than the range of using ultrasound [45]. 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 VIII-G. Moreover, different hardware can capture vocal folds vibration from 7 m in [17], or at 50 m [85]. We believe RADIOSES can benefit from better hardware significantly, and a more practical system can be built.

Latency and Complexity: As mentioned previously, RADIOSES requires a moderate amount of delay with a lookahead of 150 ms to operate causally. This delay is dominated by the delay due to the structure of the neural network, and suitable modifications that replace the dual-path structure with other efficient models can be explored to reduce the delay and computational complexity further.
**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.

**Performance Improvements and Other Applications:** Although RADIOSES presents remarkable performance improvements over audio-only systems, its performance can be further improved by incorporating different types of noise, capturing data from additional users, and training the system in a more generalized setting including simulating reverberant environments. These approaches would not only improve the performance of the system, but also would improve robustness, and make the system more applicable in real-world scenarios.

**Power Consumption and Cost:** Although our evaluation board costs $300, a single mmWave device can be purchased for $15 from TI. The transmission power of the device is 12 dBm (≈ 16 mW) and the selection of radar parameters results in a duty cycle of 7.3%, (i.e. ≈ 1.2 mW). For comparison, the size of these devices can go as small as 6 mm x 6 mm to fit in a phone [12], and the power consumption of the radar in that phone is 1 mW [12]. Furthermore, RADIOSES does not require capturing the entire signal duration (Section VIII-F) 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.

**X. CONCLUSION**

We present 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 computational complexity and privacy preservation. 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.

**REFERENCES**

[1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] G. Wichern et al., “Wham!: Extending speech separation to noisy environments,” in *Proc. InterSpeech*, 2019, pp. 1368–1372.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE Int. Conf. Acoust.*, *Speech Signal Process.*, 2016, pp. 31–35.

[4] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *Proc. IEEE Int. Conf. Acoust.*, *Speech Signal Process.*, 2017, pp. 241–245.

[5] A. Ephrat et al., “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, Jul. 2018.

[6] T. Afoureas, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” in *Proc. InterSpeech*, 2018, pp. 3244–3248.

[7] E. Z. Golumbic, B. G. Cogan, C. E. Schroeder, and D. Poeppel, “Visual input enhances selective speech envelope tracking in auditory cortex at a ‘cocktail party’”, *J. Neurosci.*, vol. 33, no. 4, pp. 1417–1426, 2013.

[8] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, “FaceFilter: Audio-visual speech separation using still images,” in *Proc. InterSpeech*, 2020, pp. 3481–3485.

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

[10] 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 Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1066–1078, Jun. 2016.

[11] “Contactless sleep sensing in nest hub with soli,” 2021. [Online]. Available: https://ai.googleblog.com/2021/03/contactless-sleep-sensing-in-nest-hub.html

[12] “Soli radar-based perception and interaction in pixel 4,” 2020. [Online]. Available: https://ai.googleblog.com/2020/03/soli-radar-based-perception-and.html

[13] “The next amazon echo could use radar to track your sleep,” 2021. [Online]. Available: https://www.techradar.com/news/the-next-amazon-echo-could-use-radar-to-track-your-sleep

[14] 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 Things J.*, vol. 8, no. 3, pp. 1294–1307, Feb. 2021.

[15] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, “Sound recovery from radio signals,” in *Proc. IEEE Int. Conf. Acoust.*, *Speech Signal Process.*, 2021, pp. 8022–8026.

[16] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, “mmVib: Micrometer-level vibration measurement with mmwave radar,” in *Proc. ACM Mobile Commun.*, 2020, pp. 1–13.

[17] Y. Luo, Z. Chen, and T. Yoshioka, “Detection of the vibration signal from human vocal folds using a 94-GHz millimeter-wave Radar,” *MDPI Sensors*, vol. 17, Mar. 2017, Art. no. 543.

[18] C. Xu et al., “Waveear: Exploring a mmWave-based noise-resistant speech sensing for voice-user interface,” in *Proc. ACM Mobile Syst.*, 2019, pp. 14–26.

[19] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, “RadioMic: Sound sensing via radio signals,” *IEEE Internet Things J.*, vol. 10, no. 5, pp. 4431–4448, Mar. 2023.

[20] T. Liu et al., “Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals,” in *Proc. 19th ACM Sensor Syst.*, 2021, pp. 97–110.

[21] 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*, vol. 17, Mar. 2017, Art. no. 543.

[22] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *Proc. IEEE Int. Conf. Acoust.*, *Speech Signal Process.*, 2017, pp. 241–245.

[23] G. Hu and D. Wang, “Speech segregation based on pitch tracking and amplitude modulation,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 79–82.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, “Monitoring vital signs using millimeter wave,” in Proc. 17th ACM MobiCom, 2016, pp. 211–220.

Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M.B. Srivastava, “Uheur: Through-wall extraction and separation of audio vibrations using wireless signals,” in Proc. ACM Sensor Syst., 2020, pp. 1–14.

Y. Dong and Y.-D. Yao, “Secure mmwave-radar-based speaker verification for IoT smartphone,” IEEE Internet Things J., vol. 8, no. 5, pp. 3500–3511, Mar. 2021.

H. Li et al., “Vocalprint: Exploring a resilient and secure voice authentication via mmwave biometric interrogation,” in Proc. ACM Sensor Syst., 2020, pp. 312–325.

M. Z. Ozturk, C. Wu, B. Wang, M. Wu, and K. J. R. Liu, “Beyond microphone: MmWave-based biometrics for robust voice activity detection,” in 1st ACM Intell. Acoust. Syst. Appl., 2022, pp. 7–12.

A. G. Stove, “Linear FMCW radar techniques,” in IEEE Proc. F (Radar Signal Process.), vol. 139, no. 5, pp. 343–350, 1992.

B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” IEEE ASSP Mag., vol. 5, no. 2, pp. 4–24, Apr. 1988.

M. A. Richards, Fundamentals of Radar Signal Processing (Professional Engineering Series). New York, NY, USA: McGraw-Hill Education, 2005.

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 Things J., vol. 8, no. 22, pp. 16623–16636, Nov. 2021.

F. Adib, Z. Kabelac, and D. Katabi, “Multi-person localization via RF shadowing,” in Proc. ISMWS, 2022, pp. 279–280.

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. 2nd Int. Conf. Knowl. Discov. Data Mining, 1996, pp. 226–231.

J. Munkres, “Algorithms for the assignment and transportation problems,” J. Soc. Ind. Appl. Math., vol. 5, no. 1, pp. 32–38, 1957.

T. Zheng, Z. Chen, S. Zhang, C. Cai, and J. Luo, “MoKe-Fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar,” in Proc. ACM Sensor Syst., 2021, pp. 111–124.

D. Michelsanti et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, no. 3, pp. 1368–1396, Mar. 2021.

N. Roman, D. Wang, and G. Brown, “Speech segmentation based on sound localization,” in Proc. Int. Joint Conf. Neural Netw., 2001, vol. 4, pp. 2861–2866.

J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 626–630.

J. S. Garofolo et al., “TIMIT acoustic-phonetic continuous speech corpus,” 1993. [Online]. Available: https://hlt-handle.net/11272/1/AB2/SWVENO

J. Cosentino, M. Pariete, S. Cornel, A. Deleforge, and E. Vincent, “Librimix: An open-source dataset for generalizable speech separation,” 2020, arXiv:2005.11262.

N. Roman, D. Wang, and D. Old, “Speech segmentation using automatic speech recognition tools,” in Proc. Interspeech 2005, pp. 2657–2660.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in Proc. Int. Conf. Mach. Learn., 2011, pp. 689–696.

E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

C. H. Tsal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2010, pp. 4214–4217.

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,” J. Audio Eng. Soc., vol. 50, no. 10, pp. 755–764, 2002.

E. Tzinis, Z. Wang, and P. Smaragdis, “Sudo RM-RF: Efficient networks for universal audio source separation;” in Proc. IEEE 30th Int. Workshop on Machine Learning for Signal Processing, 2020, pp. 1–6.

S. Vesa, “Binaural sound source distance learning in rooms,” IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

M. Yiwee and E. J. Rhee, “Sound source distance estimation using deep learning: An image classification approach,” Sensors, vol. 20, no. 1, 2020, Art. no. 172.

T. Toufours, I. S. Chung, and A. Zisserman, “My lips are concealed: Audio-visual speech enhancement through obstructions,” in Proc. InterSpeech, 2019, pp. 4295–4299.
Muhammed Zahid Ozturk (Student Member, IEEE) received the B.S. degree from the Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Türkiye, in 2018, and the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA, in 2021 and 2022, respectively. He is currently a Principal Scientist with Origin Wireless Inc., Greenbelt, MD, USA. His research interests include wireless sensing, signal processing and deep learning.

Chenshu Wu (Senior Member, IEEE) received the B.E. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2010 and 2015, respectively. He is currently an Assistant Professor with the Department of Computer Science, The University of Hong Kong, Hong Kong. He is also the Chief Scientist at Origin Wireless Inc., Greenbelt, MD, USA. His research interests include wireless AoA systems at the intersection of wireless sensing, ubiquitous computing, digital health, and the Internet of Things.

Min Wu (Fellow, IEEE) received the B.E. degree in automation and the B.A. degree in economics (Highest Hons.) from Tsinghua University, Beijing, China, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2001. She is currently a Professor and an Associate Dean of engineering and a Distinguished Scholar-Teacher with the University of Maryland, College Park, MD, USA. Her research interests include information security and forensics, multimedia signal processing, and applications of data science and machine learning in health and the IoT. She is a Fellow of AAAS and the U.S. National Academy of Inventors. She was the recipient of the U.S. NSF CAREER Award, U.S. ONR Young Investigator Award, TR100 Young Innovator Award from the MIT Technology Review, IEEE Harriet B. Rigas Education Award, and IEEE SP Society Meritorious Service Award. She was the Chair of the IEEE Technical Committee on Information Forensics and Security. She was the Vice President-Finance of the IEEE Signal Processing Society and the Editor-in-Chief of the IEEE SIGNAL PROCESSING MAGAZINE. She is currently the 2022–2023 President-Elect of the IEEE Signal Processing Society.

K. J. Ray Liu (Fellow, IEEE) is currently the Founder and President of Origin AI that pioneers AI for wireless sensing and indoor tracking. The invention of wireless AI won three prestigious CES Innovation Awards, including CES Best of Innovation in 2021. He was Distinguished University Professor, Distinguished Scholar-Teacher, and Christine Kim Eminent Professor of Information Technology of the University of Maryland, College Park, MD, USA, from where he retired after more than three decades of career in education. His research contributions encompass broad aspects of signal processing and communications. He has trained more than 70 doctoral/postdoctoral students, of which 10 are now IEEE Fellows. Dr. Liu was the recipient of two IEEE Technical Field Awards which include the 2021 IEEE Fourier for Signal Processing and 2016 IEEE Leon K. Kirchmayer Graduate Teaching Award. He was also the recipient of the IEEE Signal Processing Society 2014 Norbert Wiener Society Award and 2009 Claude Shannon-Harry Nyquist Technical Achievement Award. He recognized as a Web of Science Highly Cited Researcher. He is a Fellow of the American Association for the Advancement of Science and the U.S. National Academy of Inventors. Dr. Liu is 2022 IEEE President and CEO. He was the 2019 IEEE Vice President for Technical Activities. He was also the 2012–2013 President of IEEE Signal Processing Society, where he once was the Editor-in-Chief of IEEE SIGNAL PROCESSING MAGAZINE.