TOWARD MMWAVE-BASED SOUND ENHANCEMENT AND SEPARATION

Muhammed Zahid Ozturk[⋆†], *Chenshu Wu*[‡†], *Beibei Wang*[†], *K. J. Ray Liu*[⋆†]

[⋆]University of Maryland, College Park, MD 20742, USA
[†]Origin Wireless, Inc., Greenbelt, MD 20770, USA
[‡]University of Hong Kong, Hong Kong, China

ABSTRACT

Speech enhancement and separation have been a long-standing problem with recent advances using a single microphone. With the help of video modality, improvements have been shown for these tasks. In this work, we explore a multimodal approach using mmWave radio devices, as these devices can measure vocal folds vibration. Thorough data collection and extensive experiments with two different neural networks indicate that radio modality can bring significant improvements in speech enhancement and separation.

***Index Terms*—** mmWave sensing, sound enhancement, sound separation, deep learning

1. INTRODUCTION

Humans are enormously capable of focusing on a speaker under severe noise conditions (*a.k.a* speech enhancement (SE)) or separating one speaker from another (*a.k.a* speech separation (SS)), which is known as the cocktail party problem. Achieving high performance in these tasks using microphones and other modalities has been an interesting research problem for many years. To that end, audio-only (AO) and multimodal (*e.g.* audiovisual) methods have been studied extensively.

Audio only SE and SS methods achieved remarkable performance in the recent years with the help of deep learning. These methods usually require estimation of the number of sources, and a robust source association method. Although these problems can be solved by clustering-based methods [1] and permutation invariant training [2], audio-only methods do not exploit the environmental information fully, which we humans utilize when focusing on a speech.

To overcome those issues and to improve the performance by exploiting secondary information, multimodal systems have been investigated, such as audiovisual [3, 4]. Similar to human perception, audiovisual systems can exploit facial and lip motion and shown to improve performance when compared with audio-only baselines. On the other hand, audiovisual systems have complex processing pipelines, as the two modalities are inherently different. They require good lighting conditions and raise potential privacy concerns.

In recent years, wireless sensing [5] has been an emerging field. Among different wireless bands, mmWave-based sensing enabled many applications related to motion and vibration sensing, such as heart rate monitoring [6], measuring machinery and object vibration [7, 8], or extracting vocal folds vibration [9]. These devices can operate in dark, through-wall settings and do not raise privacy concerns. The limited information from radars [1] have been used to estimate pitch and detect voice activity [9], and reconstruct speech to some extent [10].

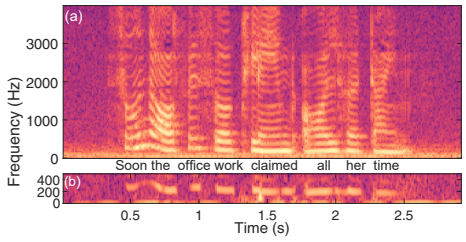

Fig. 1. (a) Audio spectrogram of a clean speech signal sampled at 8 kHz, b) Radio spectrogram of the same signal, captured from vocal folds vibration, sampled at 1kHz

Nowadays, many devices are equipped with UWB [11] or mmWave radar (e.g. Google Soli [12]), along with microphones. For SE and SS tasks, additional information about the speaker is shown to improve the performance, such as the picture of the face that gives the speaker characteristics [13], pitch estimation [14], and voice activity detection aided separation [15]. As the radio reflections from vocal folds can provide i) accurate pitch estimation, ii) detection, and iii) localization of source, and resemble a low-pass filtered version of the speech signal (as illustrated in Fig. 1), we pose the question, *how and to what extent, radio channel could contribute to the SE and SS tasks*?

In this work, we investigate the above problem by using a data collection procedure that results in both radio and audio signals, and by multimodal deep learning. Namely, our contributions are as follows:

- We collect and extract vocal folds data using an mmWave radio device, together with a microphone and camera, and convert the radar data into vibration signals.

[1]We use *radar* and *radio* interchangeably, as it is expected for radio communication devices to support radar mode of operation in the future.

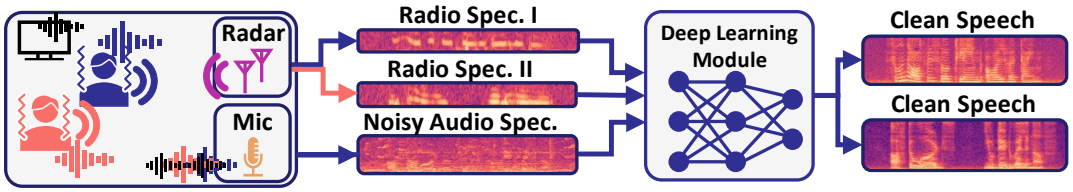

Fig. 2. System Overview

- We show, as a proof-of-concept, usefulness of the radio channel by implementing two different neural networks that combine the two modalities at different stages, and evaluate their performance extensively for speech enhancement and separation tasks.

The remainder of the paper is organized as follows. In §2, we elaborate on the system, dataset, preprocessing steps, and the deep learning module. We present and discuss our results in §3, and conclude the paper in §4.

2. AUDIORADIO MODEL

Our proposed system consists of a single microphone, accompanied by a radio device, as illustrated in Fig. 2. In a noisy environment, the microphone captures a mixture of speech and noise, whereas the radio is able to measure the vocal fold vibration of speakers separately. In order to capture vocal folds vibration signals with a radio device, we utilize the approach in our previous work [7], which extracts a real-valued time-series data that correlates with the displacement of the vocal fold's vibration, and contains the first few harmonics of the generated speech. This approach models the vibration on object surfaces and provides a method to extract the sound signal. On the other hand, the audio channel is assumed to be monaural. The system captures both modalities in sync and processes the mixed audio signal with multiple separate radio streams that are extracted from each user. The deep learning module fuses two modalities and extracts the speech of each user, which will be explained next.

2.1. Dataset

Hardware: We build a data collection platform, as seen in Fig. 3, to obtain large-scale data to train, validate, and evaluate the proposed multimodal system. We collect clean audio data with a Blue Snowball iCE microphone, sampled at 48 kHz, radar data using a Texas Instruments 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.

User setting: We recruit 18 users including native speakers and speakers with different accents to read phonetically rich sentences from the TIMIT corpus [16]. 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 user. In total, 2000 different sentences, and 5762 unique words are read by participants. During data collection, we ask users to sit approximately 40cm away from the radio device, and read each material at a normal speaking volume while remaining stationary.

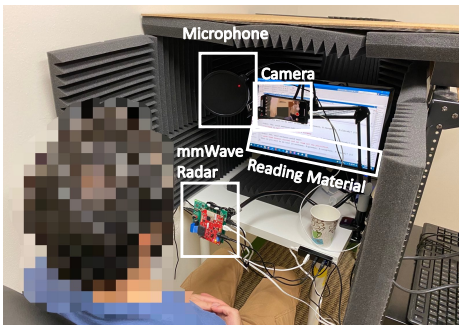

Fig. 3. Data Collection Environment

Synthetic data generation: To generate the noisy and mixture sound signals, we follow the recipe used in LibriMix [17] with the same noise files and mixing parameters. After randomly selecting 11 users for training, and 4 users (2 male, 2 female) for evaluation, we downsample the audio files to 8kHz and create synthetic mixtures based on the shortest of the combined files for multi-user mixtures. Furthermore, we drop utterances shorter than 3 seconds. Each user's recordings are repeated five times on average, and our training set only includes mixtures from the 200 common sentences, which results in 10300 utterances for 2-person mixtures. Since our dataset is limited in size and the reflection of different users can vary significantly due to posture, body motion, and physical properties, the generalization of the system needs to be tested. Therefore, we create two different evaluation sets: i) mixtures from seen users, but unheard sentences (*i.e.*, closed condition (CC)), ii) mixtures from unseen users (*i.e.*, open condition (OC)). Both test sets consist of 1000 sentences and CC performance is expected to be higher than OC. Our mixtures also include same speaker mixtures, which is a particularly challenging case.

2.2. Pre-processing

Audio: Similar to preprocessing pipeline of [3], we compute STFT of 3-second-long input audio with Hann window of length 25 ms, hop length of 10 ms, and FFT size of 512. Furthermore, we apply power-law compression on the input

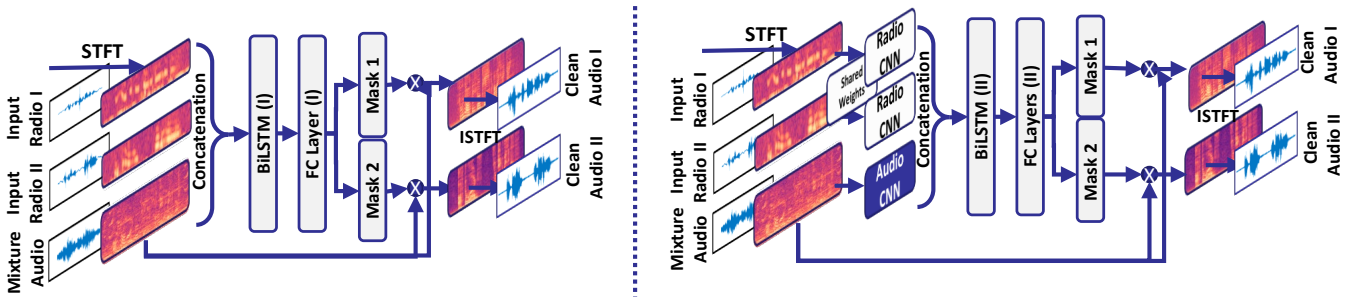

Fig. 4. Evaluated models for multiple tasks. a) Model I: Early fusion model, b) Model II: Intermediate fusion model

spectrograms, with $p = 0.3$ (*i.e.*, $S^{0.3}$, for a spectrogram S). As the inputs are 3-second segments, this procedure results in $257 \times 298 \times 2$ dimensional audio inputs.

Radio: Using the radio stream captured at 1 kHz, we calculate the radio spectrograms in a similar way to audio spectrograms, with the only difference being the FFT size of 64 points. This results in a $33 \times 298 \times 2$ dimensional representation of the radio signal. When using Model I, we resize the radio signal to 66×298 dimensions, and we apply a similar reshaping operation to the audio to make inputs compatible with the model, which will be explained next.

2.3. Deep Learning Models

We use two different models to understand changes in performance with varying models, which are illustrated in Fig. 4. The main difference between the models is the layer of *fusion* of two modalities. Both of these models use complex audio and radio spectrograms as inputs, calculated by STFT, and estimate a compressed complex ratio mask (cRM) [18] for each speaker. The output audio signals are computed by multiplying the noisy input spectrogram with the estimated masks and extracting the time domain signal by applying inverse STFT.

Model I: The first model is similar to the BLSTM based enhancement in [2], where we have a 3-layer BLSTM, followed by a single fully connected layer. This method of *early fusion* is also explored in audiovisual domain [19], but sometimes prone to mode failure, where the outputs only rely on a single modality [20]. BLSTM layer has 250 hidden layers, and outputs a $250 \times N_{\text{time}}$ feature vector, where N_{time} denotes the number of time domain samples. The output is mapped to the size of the input by a fully connected layer, which uses sigmoid function as nonlinearity to estimate a cRM.

Model II: We mimic the audiovisual model in [3] as an *intermediate-fusion* model, since the two modalities are fused after their corresponding subnetworks. To make the video subnetwork in [3] compatible with our radio inputs, we use a 5-layer CNN, with the first four layers having 128 filters with 3×3 sized kernels. The last layer has 256 kernels with varying dilation sizes. Dilation size for layer i is given as $(n_i \times n_i)$, where $n = [1, 1, 2, 4, 8]$. The same weights are shared for both streams of radio when there are multiple speakers. On the other hand, since our dataset is limited, instead of using all 15 layers in the original work [3] for audio, we use the first 8 layers before the fusion layer. The remaining fully connected layers are used without any change, and we refer the reader to the original work [3].

Training: To train these models, we use scale-invariant signal-to-distortion (SI-SDR [21]) as the loss function between the time-domain signals. A separate model for different numbers of users and tasks has been trained. In a practical scenario, the radio device can estimate the number of users in the environment and switch to the appropriate model.

2.4. Implementation and Evaluation

Implementation: We implement both models in PyTorch, using Asteroid library [22]. We use Adam optimizer with a starting learning rate of 5e-4, and reduce the learning rate by half if the validation loss does not improve for 10 epochs. Furthermore, we halve the learning rate every 20 epochs for Model I, and every 5 epochs for Model II. We train each system (*e.g.*audio baseline, audio + radio) and each task by 200 epochs using Model I, with a batch size of 32. Since Model I is relatively simple, we can train it in 3 hours using a single NVIDIA RTX 2080S GPU for 200 epochs. On the other hand, training Model II takes longer, as it has many more layers, and we train it for 40 epochs with a batch size of 6. We note that the performance could be improved further by hyperparameter search and/or with larger datasets.

Evaluation: We evaluate the performance of the SS and SE using metrics related to both tasks. Specifically, we report SI-SDR [21], PESQ [23], STOI [24] for all tasks, and signal to interference ratio (SIR) for SS, where higher values are better for all metrics. We present our evaluation in three settings, enhancing single speaker speech signals (Task 1), separating clean 2-speaker mixtures (Task 2), and separating noisy 2-speaker mixtures (Task 3).

3. RESULTS

In this section, we present the results with the proposed neural networks in Table 1 and Table 2. In both tables, A and R stand for audio and radio, whereas O stands for only (*e.g.* AO means audio-only baseline).

Table 1. Results using Model I

		Task 1: Enhance Noisy (SE)			Task 2: Separate Clean (SS)				Task 3: Separate Noisy (SS)			
	Model	SI-SDR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ
CC	Input	1.3	0.63	1.49	0.2	0.8	0.68	1.66	-3.3	0.6	0.54	1.38
	AO	9.8	0.78	1.99	**9.2**	**15.5**	**0.82**	**2.23**	4.7	14.8	0.67	1.61
	A + R	**10.1**	**0.80**	**2.02**	9.1	15.3	0.82	2.17	**6.6**	**17.6**	**0.72**	**1.68**
OC	Input	-0.2	0.59	1.44	0.1	0.5	0.67	1.62	-3.6	0.5	0.52	1.36
	AO	7.5	0.73	1.71	6.6	11.9	0.76	**1.92**	2.4	11.9	0.60	1.45
	A + R	**8.0**	**0.75**	**1.73**	**6.7**	**12.1**	**0.76**	1.91	**4.8**	**15.4**	**0.67**	**1.54**

Table 2. Results using Model II

		Task 1: Enhance Noisy (SE)			Task 2: Separate Clean (SS)				Task 3: Separate Noisy (SS)			
	Model	SI-SDR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ	SI-SDR	SIR	STOI	PESQ
CC	Input	1.3	0.63	1.49	0.2	0.8	0.68	1.66	-3.3	0.6	0.54	1.38
	AO	10.1	0.78	2.06	8.3	14.4	0.80	2.16	3.6	13.1	0.64	1.55
	A + R	**10.4**	**0.80**	**2.10**	**10.5**	**17.1**	**0.85**	**2.39**	**6.4**	**17.2**	**0.72**	**1.66**
	RO	5.2	0.67	1.48	6.6	13.0	0.77	1.82	3.3	14.4	0.65	1.43
OC	Input	-0.2	0.59	1.44	0.1	0.5	0.67	1.62	-3.6	0.5	0.52	1.36
	AO	8.3	0.74	1.82	4.6	10.0	0.70	1.71	1.5	10.2	0.56	1.39
	A + R	**8.9**	**0.76**	**1.87**	**8.3**	**14.3**	**0.80**	**2.06**	**5.0**	**15.3**	**0.67**	**1.51**
	RO	4.3	0.65	1.40	5.8	11.7	0.75	1.74	2.6	13.5	0.63	1.38

Model I: As presented in Table 1, a radio-based system can bring improvements to an audio-only baseline, even with the early fusion of the two modalities. The addition of radio improves all metrics in noisy tasks, and the highest amount of improvement is observed when separating noisy mixtures. We do not observe much difference for clean mixtures, which we had mode failure as the only audio signal dominated the neural network. On the other hand, noisy SS task benefits the most from our multimodal system, and it generalizes better to unseen user cases, compared to audio-only baseline. We observe a relatively large difference between closed condition and open condition tasks when compared to the other papers in the literature. However, the decline in the multimodal system is lower compared to the audio-only case. We believe, by using a larger dataset, the performance degradation could be mitigated for all systems, and we leave this to future work.

Model II: In this model, we observe that the gap between audio-only and radio-based speech enhancement is larger for SS tasks. Here, we also include results from radio-only training, which only uses the radio signal to estimate the mask. Similar to results with Model I, greater improvements can be achieved in SS tasks, in contrast to single speaker enhancement. Furthermore, even a radio-only system outperforms audio-only methods for SS tasks, which indicates the usefulness of the radio channel alone. The performance with respect to perceptual quality is lower, which is usually an issue with spectrogram representations [20].

4. CONCLUSION AND DISCUSSION

In this work, we explore the benefits of having the radio channel data for speech enhancement and separation. Our results indicate that radio modality can complement the microphone and can bring improvements to these tasks even with a simple neural network. Significant improvements can be achieved especially for speech separation tasks. Our results indicate that many interesting problems in SE and SS can be mitigated by the radio channel considerably, which can be a great alternative to the video channel, with the advantages of being robust against changes in lighting, and being privacy preserving. Despite promising results, there is a lot of additional work and future directions that need to explore the full potential of the radio channel. These include:

Multi-user evaluation: Appropriate signal processing algorithms to detect/localize (multi)users, and evaluate the performance with real-mixtures (*in-the-wild*) is needed. Since our mixtures assume clean radar signals, the performance can potentially decrease with nearby users and increasing distance, and the system needs to be verified in challenging conditions, with higher amounts of noise and distortion.

Better models: The radio channel information is not necessarily similar to that of the video and more suitable deep learning models need to be explored. Our objective was not to find the optimal structure, and we leave this to future work.

Other side channels: Radars can not only measure vocal folds vibration, but also vibration of other sources, such as guitars [25], or machinery vibration [8], and these radar signatures can be used to improve quality in SE and SS tasks.

Microphone arrays: Radars can also complement microphone arrays by estimating the direction of sources robustly in noisy environments, which is a challenging problem [20].

Other Tasks: Similarly, radars can also be applied to musical source separation, or dereverberation tasks with suitable models.

5. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE ICASSP*, 2016, pp. 31–35.

[2] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE ICASSP*, 2017, pp. 241–245.

[3] 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, no. 4, Jul. 2018.

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

[5] 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.

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

[7] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "Sound recovery from radio signals," in *IEEE ICASSP*, 2021, pp. 8022–8026.

[8] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmVib: Micrometer-level vibration measurement with mmWave radar," in *ACM MobiCom*, 2020, pp. 1–13.

[9] 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.

[10] 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 *ACM MobiSys*, 2019, pp. 14–26.

[11] "Apple airtags use uwb wireless tech." https://www.cnet.com/tech/mobile/apple-airtags-use-uwb-wireless-tech-heres-how-ultra-wideband-makes-your-life-easier/, accessed: 2021-09-23.

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

[13] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "FaceFilter: Audio-visual speech separation using still images," in *Interspeech*, 2020, pp. 3481–3485.

[14] 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 TASLP*, vol. 24, no. 6, pp. 1066–1078, 2016.

[15] 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.

[16] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993. [Online]. Available: https://hdl.handle.net/11272.1/AB2/SWVENO

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

[18] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, no. 3, pp. 483–492, 2016.

[19] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *IEEE ICASSP*, 2019, pp. 6900–6904.

[20] 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 TASLP*, vol. 29, pp. 1368–1396, 2021.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE ICASSP*, 2019, pp. 626–630.

[22] 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 *Interspeech*, 2020, pp. 2637–2641.

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE ICASSP*, 2001, pp. 749–752.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE ICASSP*, 2010, pp. 4214–4217.

[25] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "RadioMic: Sound sensing via mmWave signals," *CoRR*, vol. abs/2108.03164, 2021. [Online]. Available: https://arxiv.org/abs/2108.03164