



Radio2Speech: High Quality Speech Recovery from Radio Frequency Signals

Running Zhao^{1,3}, Jiangtao Yu^{2,3}, Tingle Li^{2,3}, Hang Zhao^{2,3,*}, Edith C.H. Ngai^{1,*}

¹The University of Hong Kong, Hong Kong SAR, China

²IIS, Tsinghua University, Beijing, China

³Shanghai Qi Zhi Institute, Shanghai, China

hangzhao@mail.tsinghua.edu.cn, chngai@eee.hku.hk

Abstract

Considering the microphone is easily affected by noise and soundproof materials, the radio frequency (RF) signal is a promising candidate to recover audio as it is immune to noise and can traverse many soundproof objects. In this paper, we introduce Radio2Speech, a system that uses RF signals to recover high quality speech from the loudspeaker. Radio2Speech can recover speech comparable to the quality of the microphone, advancing from recovering only single tone music or incomprehensible speech in existing approaches. We use *Radio UNet* to accurately recover speech in time-frequency domain from RF signals with limited frequency band. Also, we incorporate the neural vocoder to synthesize the speech waveform from the estimated time-frequency representation without using the contaminated phase. Quantitative and qualitative evaluations show that in quiet, noisy and soundproof scenarios, Radio2Speech achieves state-of-the-art performance and is on par with the microphone that works in quiet scenarios.

Index Terms: speech recovery, radio frequency, wireless sensing

1. Introduction

The microphone “listens” to ambient audio like the human ear, giving rise to application in human-computer interaction and security (e.g., eavesdropping). However, microphones only perceive the audio and does not possess the sensitivity and discrimination of the human ear, so it is less capable of resisting noise and irrelevant speech [1, 2, 3]. Further, both the human ear and microphone will be “deaf” when there are soundproof materials between the audio source and receiver. The instinct deficiencies of microphones hinder their use in noisy and soundproof scenarios. Recent researches have proposed to use other modalities of microphones that allow them to sidestep these shortcomings, such as laser [4], Lidar [5], and visual microphones [6]. Yet, they are susceptible to lighting conditions or opaque substances.

Radio frequency (RF) based sensing systems are possible candidates to deal with all these issues in a holistic way. This is based on the fact that RF signals are insensitive to surrounding acoustic noise and lighting and can traverse occlusions [7, 8]. Recent advances in RF-based systems have leveraged those properties to recognize or recover audio. Preliminary works [9, 10, 11] leveraged RF signals to recognize several words or phonemes according to the movement of vocal organ. Furthermore, in [12], a RF eavesdropping system was proposed to recover audio from the loudspeaker. UWHeard [13] presented a system to recover and separate sounds from multiple loudspeakers based on RF signals. However, these systems can only recover the single tone music and speech with simple sentences (e.g., “one two three”). It is conceivable that speech (please note

the speech here and later refers to speech in the spoken corpus) contains complex frequency components and irregular harmonics compared with them, and thus speech recovery is a difficult task. Recently, WaveEar [14] tried to recover the human reading voice via radar. Although it achieves good quantitative results, the recovered speech is incomprehensible to listeners. RadioMic [15] used RF signals to recover speech from the loudspeaker and medium, but it suffers from the same problem. The core function of microphones is to record the audio as close to the original as possible so that it is intelligible to listeners. This principle should also be followed by the RF microphone even though it has unique advantages in noisy and soundproof scenarios. Therefore, using RF signals to recover high quality speech has remained intractable.

In this paper, we introduce Radio2Speech, a system that uses RF signals to recover speech of corpus with high quality from the loudspeaker. Nowadays, loudspeakers are ubiquitous in meeting room and home theater as a common audio source. Thus, it makes sense to acquire information from speech of loudspeakers. As illustrated in Figure 1, our system transmits the mmWave frequency-modulated continuous wave (FMCW) signal to the loudspeaker, and parses the reflected signal for recovery. In quiet scenario, Radio2Speech can recover high quality speech like a microphone. Even though in noisy and soundproof scenarios where microphone fails, our system is still able to recover high quality speech like in quiet scenario.

The distinctive properties of RF signals constrain the high quality speech recovery. First, restricted to device design, FMCW radar is not as sensitive as the diaphragm of microphones, and the upper frequency that RF signals can perceive is limited (below 1KHz). Also, its sampling rate is only 5.1KHz that is far lower than that of speech signals. Therefore, the high frequency band of RF signals is missing, which directly influences the speech quality. Second, RF signals contain much noise from FMCW radar hardware itself (referred to as phase noise [16]) and environmental noise due to the multipath effect. Such noise contaminates the frequency and phase of the estimated speech signal. The above properties of RF signals have implications for speech quality, and proper handling of these properties is the key to recovering high quality speech.

Radio2Speech takes into account the above properties and deals with them as follows. Based on the UNet style network, we incorporate the frequency transformation layer to exploit correlation among all frequency harmonics, which have been proven to be helpful for time-frequency representation (TFR) reconstruction [17]. This incorporation allows the network to make full use of the limited frequency information to predict missing high frequency band due to limited frequency perception range and low sampling rate. Our UNet style backbone, named TUNet, employs Transformer layers at the bottleneck of convolutional neural network (CNN) based UNet due to its

*Corresponding Author

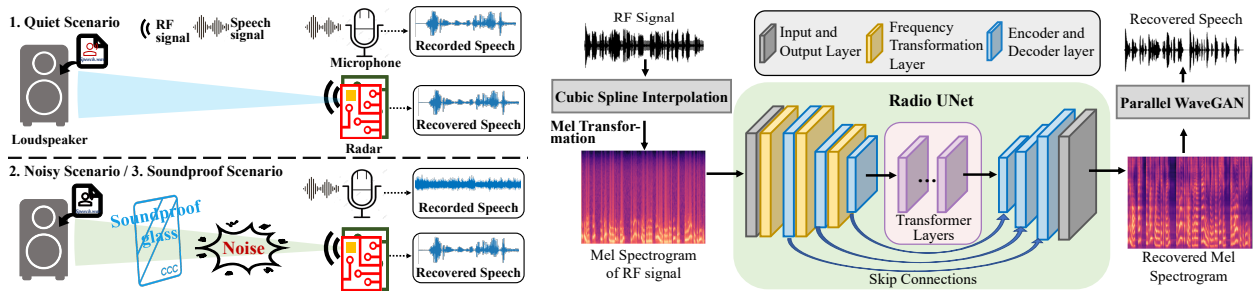


Figure 1: Illustration of Radio2Speech working scenarios (Left). In quiet scenario, Radio2Speech recovers the speech like the microphone. In noisy and soundproof scenarios, it performs the same as before, while the microphone fails. The architecture of Radio2Speech (Right). The input RF signal is upsampled to 8KHz and the Mel spectrogram of the RF signal is used as the input of Radio UNet.

effectiveness in learning the contextual information in time-frequency domain. The combination of them constitutes *Radio UNet*. Beyond this, we introduce a neural vocoder to synthesize speech signals from the recovered TFR, preventing the use of contaminated phase for synthesis. Experimental results show that in quiet, noisy and soundproof scenarios, Radio2Speech yields state-of-the-art quantitative and qualitative scores and is comparable to the microphone that works in quiet scenarios.

2. Radio2Speech

The general goal of Radio2Speech is to parse RF signals and recover high quality speech. Figure 1 describes the overall pipeline. First, the input RF signal is upsampled from 5.1KHz to 8KHz using the cubic spline interpolation that is normally used in bandwidth extension [18]. Then, Mel transformation is applied to the upsampled RF signal to obtain its Mel spectrogram. Next, it is fed into *Radio UNet* to recover the Mel spectrogram of the speech signal. Finally, the neural vocoder, Parallel WaveGAN, is employed to reconstruct the natural speech waveform from the estimated Mel spectrogram.

2.1. Speech Recovery in Time-Frequency Domain

We employ *Radio UNet* to achieve recovery from the Mel spectrogram of RF signals to that of speech signals. Taking into account RF signal properties, *Radio UNet* performs not only the transformation of Mel spectrograms, but also the speech bandwidth extension due to the low sampling rate and limited frequency perception range. Considering the Mel spectrogram pair of RF and speech signals (M_{RF}, M_s) fed to our network as input and target, the *Radio UNet* acts as a function $T(\cdot)$ to perform the Mel spectrogram recovery $\tilde{M}_s = T(M_{RF})$.

In contrast to prior works using CNN based UNet as backbones in the TFR prediction task, our TUNet has the ability to capture long-range dependencies and global context, which CNN does not possess due to its localized receptive fields. The Mel spectrogram has a strong correlation along the whole time and frequency axes, and thus capturing contextual information is beneficial for our task. Inspired by the success of Transformer based UNet in image segmentation [19], our TUNet employs Transformer layers at the bottleneck so that network leverages contextual information in time-frequency domain. Specifically, the feature map $x \in \mathbb{R}^{H \times W \times C}$ with spatial resolution $H \times W$ and C channels extracted from CNN encoder are directly reshaped as the sequence of flattened patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ to perform tokenization, where $N = HW/P^2$ is the number of patches, and P is set 1 here. Then, the patches are projected into D-dimensional patch embeddings by a linear projection, and a

learnable positional embedding is added to them. The resulting sequence is the input of stacked Transformer layers, followed by CNN decoder.

As mentioned earlier, the high frequency band of RF signals is missing. It is crucial to make full use of the limited frequency band of RF signals for high frequency prediction. Especially, harmonic correlation along the whole frequency axis has been proven to be helpful for time-frequency representation reconstruction [17]. Although the backbone TUNet has the ability to capture contextual correlation for prediction, it pays less attention to such harmonics. Thus, we introduce the frequency transformation block from [20], but simplify it as the frequency transformation layer (FTL) to capture harmonics from RF signals with limited frequency band. FTL consists of three stacked CNN layers, a fully connected layer used as a transformation matrix, and a CNN layer used for concatenation. Assuming the input features extracted from the stacked CNN layers at time step t as $f_{in}(t) \in \mathbb{R}^{F \times C}$ ($t = 1, \dots, T$), where T, F and C are the time (width), frequency (height) and channel dimension, and transformation matrix as $W_{tr} \in \mathbb{R}^{F \times F}$, the output features f_{out} after applying transformation matrix on input features can be represented as $f_{out}(t) = W_{tr} \cdot f_{in}(t)$. The transformed features f_{out} stacked along the time axis contain the global frequency correlations, and then they are concatenated with the input features f_{in} by a 1×1 convolution to capture both global and local frequency correlations. The FTL is added after input layer and each encoder layer. This allows the network to make full use of global frequency in limited frequency band of RF signals, constructing the Mel spectrogram with high frequency.

The L1 loss between Mel spectrogram of ground truth M_s and estimated Mel spectrogram \tilde{M}_s is used for training.

2.2. Speech Waveform Synthesis

We use the neural vocoder, Parallel WaveGAN [21], to synthesize speech waveform. The estimated Mel spectrogram of speech signal \tilde{M}_s is fed into the trained vocoder to convert the frequency acoustic features into natural sounding speech.

The use of this vocoder considers the RF signal property that the noise contaminates its phase. Phase has a significant effect on the perceptual quality of generated speech waveform [22]. The intuition is to directly use inverse Short Time Fourier Transform (iSTFT) on the phase of input spectrogram of RF signals and estimated spectrogram to recover speech. However, as mentioned earlier, the phase of RF signals is contaminated by noise, and thus there is a large gap between the phase of RF spectrograms and target speech spectrograms. Directly using such a phase for speech synthesis affects speech quality.

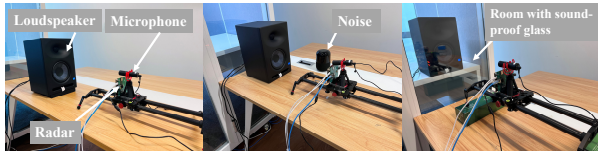


Figure 2: *Experimental scenarios. From left to right are quiet, noisy and soundproof scenarios.*

Also, there is little structure in the phase spectrogram, making it difficult to predict the phase spectrogram while estimating the magnitude spectrogram. Considering that *Radio UNet* recovers accurate Mel spectrogram with affluent speech features, synthesizing speech waveform from Mel spectrogram can effectively avoid such shortcomings. Among the many neural vocoders, we adopt Parallel WaveGAN to transform the recovered Mel spectrogram into a natural speech waveform due to its fast and small-footprint characteristics.

3. Experiments and Results

3.1. Experimental Setup

Dataset. We used two corpora, TIMIT [23] for multi-speaker scenario and LJSpeech [24] for single-speaker scenario, as the speech source for loudspeaker playback. LJSpeech consists of 13100 clips of a single speaker. The length of each clip varies from 1 to 10 seconds, and the total length is approximately 24 hours. TIMIT includes 6300 clips uttered by 630 speakers with different accents lasting 3.5 hours. The samples from such corpora are downsampled into 8KHz as the ground truth.

We used a COTS mmWave FMCW radar TI AWR1642 and a data capture board DCA1000EVM to collect RF data corresponding to the above two corpora. After calibration, the length error between original speech and RF data is less than 10 ms. We conduct RF data collection experiments from three scenarios (see Figure 2). (i) In the quiet scenario, the loudspeaker and radar were placed in a line and at a distance of 0.5m. (ii) The configuration of the noisy scenario was the same as the quiet scenario, but with an extra loudspeaker playing noise around the radar. (iii) While in the soundproof scenario, radar and loudspeaker were separated by soundproof glass, and the distance between them was 0.3m. In these scenarios, speeches were played from the loudspeaker at 85 dB SPL, similar to an actor’s stage sound [6]. Also, in these scenarios, a microphone was placed in the same location as the radar to collect data, which is used for comparison in Section 3.3. According to our investigation, the system performance decreases with increasing range (between radar and loudspeaker), similar to microphones. All because the signal (i.e., RF and audio signals) is attenuated as it propagates [25]. The system performance also degrades with deviation in angle (with respect to radar) like unidirectional microphones, due to the limited perception angle [26].

For the data from LJSpeech, the last 1000 samples were reserved for model evaluation and the remaining were used for training. Moreover, we randomly selected one sample from each speaker of the TIMIT, a total of 630 samples, for evaluation, and the others were used for training.

Pre-processing. Input RF signals were upsampled from 5.1KHz to 8KHz and target speech signals were downsampled to 8KHz before feeding into our network. Then, we extracted 80-dimensional log-Mel-spectrogram features with band-limited frequency range (60 to 4000 Hz). The window and hop sizes were set to 512 and 128.

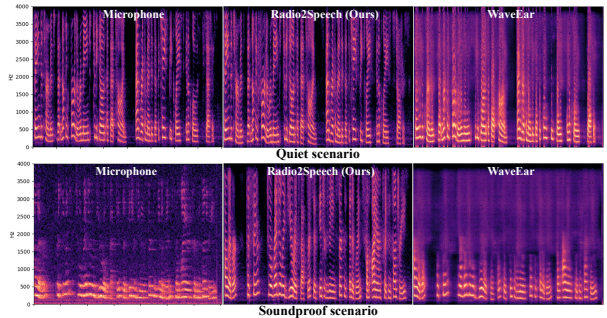


Figure 3: *Spectrograms of recorded speech and recovered speech. The top and bottom rows correspond to quiet and soundproof scenarios, respectively.*

Implementation details. TUNet has 3 encoder and decoder layers, 12 Transformer layers, and 1 input layer and output layer. All kernel sizes of CNN are set to 3×3 . Also, CNN with stride size of 2×2 in encoder is used for downsampling, while decoder uses pixel shuffle to upsample feature maps. The configurations of FTL follow the settings in [20]. The Transformer layers were pretrained on ImageNet [27], and the Parallel WaveGAN were trained by the samples with 8KHz of the corresponding corpus. We use the normalized log-Mel-spectrogram with a fixed size of $80 \times 80 \times 1$ ($T \times F \times C$, about 1.28s) as the input for training efficiency. Our network was trained using the SGD optimizer with a learning rate of 0.01.

3.2. Evaluation Metrics

We use three quantitative metrics to evaluate the quality of recovered speech. **Short-time objective intelligibility (STOI)** measures the speech intelligibility (from 0 to 1). **Perceptual evaluation of speech quality (PESQ)** for narrow band is a perception evaluation related to subjective opinion (from 1 to 4.5). Signal-to-noise ratio (SNR) is frequently used to evaluate reconstructed speech quality. However, the SNR value of speech synthesized from the generative model is poor as its output is not exactly aligned with the target at sample level [28]. In this case, SNR may not truly reflect the speech quality, thus, we choose **log-spectral distance (LSD)** that measures quality in frequency domain as the metric, and relevant parameters follow the settings in [29]. All metrics except LSD are better if higher.

We also use **mean opinion score (MOS)** to perform the subjective evaluation on Amazon Mechanical Turk. For each scenario, we randomly selected 10 RF data from the testing set of LJSpeech and TIMIT, respectively, a total of 20, and got the recovered speech from Radio2Speech and WaveEar. We also provided the corresponding speech recorded from the microphone. Each utterance was evaluated by 20 native speakers and the listeners were asked to rate a score from 1 to 5 (higher score means better quality) on speech naturalness.

3.3. Speech Recovery Results

We compare Radio2Speech with microphone and WaveEar [14] in quantitative and qualitative evaluations. Some examples are available online¹. Although WaveEar is designed to recover speech based on throat vibration, its network is also suitable for loudspeaker speech recovery and it outperforms other relevant methods. Thus, we re-implemented WaveEar consisting of encoder-decoder convolutional network and GriffinLim al-

¹<https://zhaorunning.github.io/Radio2Speech/>

Table 1: Quantitative evaluation results of ours, reference microphone (Mic) and WaveEar [14] in three experimental scenarios with respect to LJSpeech and TIMIT. The best results are shown in bold. Since RF signals are not affected by the noise in our noisy scenario, the metrics are similar in noisy and quiet scenarios.

Dataset	Metrics	Quiet scenario			Noisy scenario			Soundproof scenario		
		Ours	Mic	WaveEar [14]	Ours	Mic	WaveEar [14]	Ours	Mic	WaveEar [14]
LJSpeech	LSD↓	0.90	1.05	1.27	0.90	2.56	1.27	1.02	3.55	1.44
	STOI↑	0.89	0.90	0.74	0.89	0.30	0.74	0.75	0.63	0.62
	PESQ↑	2.50	3.00	1.78	2.50	1.24	1.78	2.09	1.59	1.61
TIMIT	LSD↓	0.97	1.04	1.23	0.97	2.78	1.23	1.07	2.24	1.35
	STOI↑	0.80	0.92	0.68	0.80	0.29	0.68	0.70	0.59	0.58
	PESQ↑	2.00	3.25	1.63	2.00	1.26	1.63	1.71	1.60	1.43

Table 2: MOS results on speech naturalness. It is with 95% confidence intervals.

Scenario	Ours	Mic	WaveEar [14]
Quiet	3.10±0.16	3.21±0.15	2.77±0.18
Noisy	3.18±0.15	1.21±0.17	2.52±0.15
Soundproof	3.08±0.14	1.43±0.15	2.38±0.18

Table 3: Ablation study on LJSpeech collected in quiet scenario.

Method	LSD↓	STOI↑	PESQ↑
Radio2Speech (Ours)	0.90	0.89	2.50
w/o Transformer	0.93	0.86	2.37
w/o freq trans layer	0.98	0.85	2.21
w/o PWG+GriffinLim	1.04	0.82	1.94
w/o PWG+iSTFT	1.08	0.77	1.52

gorithm [30] for comparison. For fair comparison, all methods was trained and evaluated by the identical training and test sets.

Figure 3 uses spectrograms to visualize the recorded and recovered speech. Since acoustic noise does not affect RF-based systems, the recovered spectrogram in noisy scenario is similar to that in quiet scenario, and we do not show it here due to space constraints. As can be seen, in quiet scenarios, Radio2Speech recovers comparable speech to the microphone, and it still performs well even in soundproof scenarios where the microphone completely fails. Moreover, Radio2Speech outperforms WaveEar in all scenarios.

Quantitative Evaluation. Table 1 presents the quantitative results, evaluated by three metrics. It can be seen that in the quiet scenario, the metrics of Radio2Speech are almost as good as microphone for LJSpeech. Although ours performs slightly worse on the TIMIT dataset due to its limited samples and multi-speaker scenario, it is acceptable. Thus, Radio2Speech has a comparable ability with the microphone in terms of speech intelligibility and quality in the quiet scenario. Moreover, since RF signals are not affected by noise outside its narrow perception field, metrics of RF-based systems in noisy and quiet scenarios are the same. This indicates that Radio2Speech is still capable of recovering high quality speech in the noisy scenario. As for the soundproof scenario, although the existence of soundproof glass attenuates the RF signal power, metrics show that the speech recovered by Radio2Speech remains intelligible. As we expected, the microphone almost fails in noisy and soundproof scenarios, which is the instinctive advantage of RF microphones. These results demonstrate that Radio2Speech can recover speech comparable to that recorded by the microphone.

Radio2Speech outperforms WaveEar by a large margin across all three metrics for LJSpeech and TIMIT datasets in three experimental scenarios. Especially in the soundproof scenario, our metrics remain high, while those of WaveEar show its recovered speech is incomprehensible. Therefore, the intelligibility and quality of speech recovered from Radio2Speech are better than the state-of-the-art (SOTA) system, suggesting that our Radio2Speech can better exploit the properties of RF signals to recover high quality speech.

Qualitative Evaluation. Table 2 shows the qualitative results of MOS evaluation, which can better show the actual perceptual quality of the recovered speech. As can be seen, in the quiet scenario, Radio2Speech achieves competitive performance compared to a microphone. In noisy and soundproof scenarios, our MOS scores remain high, which is on par with a microphone works in the quiet scenario. It concludes that Radio2Speech can recover high quality speech like a microphone in quiet, noisy and soundproof scenarios. Moreover, Radio2Speech surpasses WaveEar in all three scenarios, manifesting our system has better speech recovery ability than the SOTA system.

3.4. Ablation Study

Table 3 reports the ablation study results, and all experiments were carried out on the testing set of LJSpeech collected in the quiet scenario. *w/o Transformer* and *w/o freq trans layer* represent that the Transformer layer and FTL are removed from Radio2Speech, respectively. All the metrics drop without Transformer, indicating the backbone of TUNet is better than UNet. By comparing *w/o freq trans layer* to ours, we find that FTL provides 0.04 and 0.29 gain on STOI and PESQ, manifesting FTL can improve the speech quality by fully exploiting harmonic. Moreover, *w/o PWG+GriffinLim* and *w/o PWG+iSTFT* show the performance of using the GriffinLim algorithm and iSTFT to construct waveform instead of Parallel WaveGAN. Both of them result in 0.14-0.18 gain on LSD, 0.07-0.12 drop on STOI and 0.56-0.98 drop on PESQ. This demonstrates that the phase contaminated by noise significantly influences speech quality and the phase estimated by GriffinLim is not accurate.

4. Conclusions

In this paper, we presented Radio2Speech, a system that parses RF signals to recover high quality speech from the loudspeaker. Experimental results show that in the quiet scenario, the performance of Radio2Speech is comparable to the microphone, and our system remains effective in noisy and soundproof scenarios where the microphone almost fails. Also, in these scenarios, Radio2Speech achieves SOTA performance in both quantitative and qualitative metrics. We believe Radio2Speech could be utilized to perform more speech downstream tasks in the future.

5. References

- [1] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2015.
- [2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.
- [3] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proceedings of European Conference on Speech Communication and Technology*, 2001.
- [4] R. P. Muscatell, "Laser microphone," *The Journal of the Acoustical Society of America*, vol. 76, no. 4, pp. 1284–1284, 1984.
- [5] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: eavesdropping via lidar sensors," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 354–367.
- [6] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, jul 2014.
- [7] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7356–7365.
- [8] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 2018, pp. 267–281.
- [9] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2014, pp. 593–604.
- [10] D. Ferreira, S. Silva, F. Curado, and A. Teixeira, "Rassper: Radar-based silent speech recognition," in *Proc. Interspeech*, vol. 2021, 2021, pp. 646–650.
- [11] P. Birkholz, S. Stone, K. Wolf, and D. Plettemeier, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2404–2411, 2018.
- [12] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2015, pp. 130–141.
- [13] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "Uwhear: through-wall extraction and separation of audio vibrations using wireless signals," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 1–14.
- [14] 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 *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2019, pp. 14–26.
- [15] M. Z. Ozturk, C. Wu, B. Wang, and K. Liu, "Radiomic: Sound sensing via mmwave signals," *arXiv preprint arXiv:2108.03164*, 2021.
- [16] K. Siddiq, M. K. Hobden, S. R. Pennock, and R. J. Watson, "Phase noise in fmcw radar systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 1, pp. 70–81, 2019.
- [17] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1559–1569, 2018.
- [18] H. Wang and D. Wang, "Time-frequency loss for cnn based speech super-resolution," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 861–865.
- [19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [20] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 05, 2020, pp. 9458–9465.
- [21] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.
- [22] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.
- [23] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.
- [24] K. Ito and L. Johnson, "The lj speech dataset," <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [25] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.
- [26] G. M. Sessler and J. E. West, "Second-order gradient unidirectional microphones utilizing an electret transducer," *The Journal of the Acoustical Society of America*, vol. 58, no. 1, pp. 273–278, 1975.
- [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009, pp. 248–255.
- [28] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "Nu-gan: High resolution neural upsampling with gan," *arXiv preprint arXiv:2010.11362*, 2020.
- [29] J. Lee and S. Han, "NU-Wave: A Diffusion Probabilistic Model for Neural Audio Upsampling," in *Proc. Interspeech*, 2021, pp. 1634–1638.
- [30] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.