ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11460763

RADARNET: RADAR-ACOUSTIC DENOISING ALIGNMENT AND RESTORATION NETWORK FOR ICASSP 2026 RASE CHALLENGE

Wei Liu[1,2]*, Xu Shen*[2]*, Yihao Meng*[1]*, Gongping Huang*[1,*]*, Shoji Makino*[2]

[1]School of Electronic Information, Wuhan University, Wuhan, Hubei 430072, China
[2]Graduate School of Information, Production and Systems, Waseda University, 808-0135, Japan

ABSTRACT

Radar-based speech acquisition offers a compelling alternative to microphones in physically occluded scenarios, but it remains challenged by heavy noise contamination and intrinsically band-limited frequency response. To address these issues, we propose a two-stage framework termed the Radar–Acoustic Denoising Alignment and Restoration Network (RADARNet) for full-band speech reconstruction from radar recordings. RADARNet consists of two complementary stages: a denoising-and-alignment module for baseband recovery and a restoration module for high-frequency reconstruction. According to the official results, RADARNet ranks 1st in the ICASSP 2026 RASE Challenge.

***Index Terms*—** ICASSP Grand Challenge, radar acoustics, speech enhancement.

1. INTRODUCTION

The ICASSP 2026 Radar Acoustic Speech Enhancement (RASE) Challenge [1] targets the recovery of intelligible, full-band speech from radar vibrometry signals. Unlike conventional microphones that measure air pressure variations, radar vibrometry acquires speech information indirectly by detecting micrometer-level displacements on vibrating surfaces, such as the human throat, loudspeakers, or nearby objects, thereby enabling non-contact sound acquisition and sensing through occluding materials like glass walls [2, 3, 4]. This capability opens up new avenues for robust acoustic sensing in adverse environments, including through-barrier communication, privacy-preserving voice interfaces, and covert acoustic monitoring [5, 6, 7].

In this work, we propose a two-stage framework, termed the Radar–Acoustic Denoising Alignment and Restoration Network (RADARNet), which explicitly addresses both the noise contamination and the bandwidth limitation of radar-acquired speech. Stage 1 uses a TF-GridNet [8] module for denoising, combined with a progressive alignment strategy to compensate for the radar–microphone temporal mismatch and recover the baseband speech structure. Stage 2 adopts a TF-Restormer [9] to refine spectral details and reconstruct the missing high-frequency components. This cascaded design allows each stage to focus on a well-defined sub-task, yielding improved robustness and perceptual quality. Our system is trained strictly under the official competition protocol without any external data or pre-trained models. Experimental results demonstrate that RADARNet achieves superior speech enhancement performance.

* Corresponding author. Wei Liu and Xu Shen made equal contributions.
This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62471340, and in part by the China Scholarship Council (CSC) under Grant 202506270089.

2. APPROACH

An overview of the RADARNet architecture is shown in Fig. 1, followed by details on the network design and training strategy.

2.1. Stage 1: Denoising and Alignment

To enhance speech from heavily corrupted radar measurements, in the first stage, we adopt TF-GridNet [8] as the denoising module and introduce a phase encoder [10] before denoising to improve phase modeling. A key challenge in radar-based speech enhancement lies in the mismatch between the observed signal and the supervision signal. Specifically, the noisy observation is acquired by a mmWave radar, while the clean reference speech is recorded by a near-field microphone. Since these two signals originate from independent devices, a misalignment issue naturally arises, which significantly hinders supervised training. In practice, the radar and microphone operate on independent clocks, resulting in a sampling rate offset and a progressive temporal drift over long recordings. Due to severe noise corruption, direct alignment between the raw noisy signal and the reference speech is unreliable.

To address the misalignment, we explicitly align the reference signal with the denoised output of Stage 1. Specifically, the relative sampling rate offset is first estimated [11] and then compensated for by resampling the reference signal to match the denoising output. After resampling, a residual time delay may still exist. This delay is estimated using the GCC-PHAT method [12]. We adopt a progressive alignment strategy during training, allowing the alignment process to become increasingly accurate as denoising improves. To stabilize training in its initial phase and provide reliable temporal cues for alignment, we introduce a decaying residual connection from the noisy signal $\mathbf{y}$ to Stage 1 output $\hat{\mathbf{s}}_1$:

$$\hat{\mathbf{s}}_1 \leftarrow \hat{\mathbf{s}}_1 + \lambda \mathbf{y}, \tag{1}$$

where λ is a learnable scalar initialized as 1 and gradually driven toward 0 during training, enabling a smooth transition from input-guided alignment to model-dominated enhancement. Let $\hat{\mathbf{Z}}_{1,\text{mag}}$ and $\mathbf{Z}_{\text{mag}}$ denote the magnitude of the spectrogram of Stage 1 output and the aligned target speech, respectively. The loss of Stage 1 is defined as a log-magnitude spectral distance

$$\mathcal{L}_{\text{S1}} = \left\| \log(1 + \hat{\mathbf{Z}}_{1,\text{mag}}) - \log(1 + \mathbf{Z}_{\text{mag}}) \right\|_{\text{F}}^2, \tag{2}$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm.

2.2. Stage 2: High-Frequency Restoration

Although Stage 1 recovers the low-frequency (baseband) structure of speech, the high-frequency components often remain attenuated

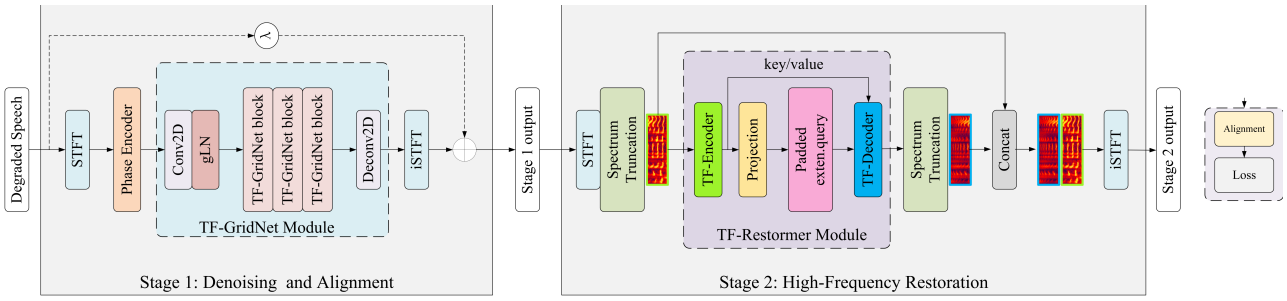

Fig. 1. Overview of the proposed RADARNet.

or missing. Stage 2 therefore adopts a TF-Restormer [9], an encoder–decoder module that analyzes the input bandwidth through a time–frequency dual-path encoder and reconstructs the missing high-frequency bands using a lightweight decoder. The restored spectral details are refined by exploiting the low-frequency information inferred in Stage 1. Specifically, the input spectrum is first truncated to retain only the low-frequency components, which are then fed into a TF-encoder. The TF-encoder, which alternately applies frequency and time modules [9], is designed to extract speech-related representations from the input signals and to model spectro-temporal dependencies. The encoder features are subsequently injected into the decoder via two pathways: (i) through channel projection and addition as a conditioning signal, and (ii) by serving as the key and value in a cross self-attention [13]. To facilitate high-frequency restoration, a set of learnable extension queries is introduced at the decoder input. The reconstructed high-frequency components are then concatenated with the low-frequency input to form the full-band spectrum. Stage 2 is trained with the same loss as Stage 1, which is computed only over the high-frequency spectrum.

3. EXPERIMENTS

3.1. Implementation

All experiments were conducted strictly on the officially provided dataset, which comprises approximately 18.7 hours of training data and 2.7 hours of validation data. The audio samples were single-channel recordings sampled at 8 kHz, and no data augmentation strategies were applied. The cutoff frequency that distinguishes between high and low frequencies was set to 1 kHz. For the time-frequency transformation, we utilized the STFT with a Hamming window, setting the window length to 32 ms and the hop size to 8 ms. For Stage 1, the output channel of the phase encoder was set to 4 and the denoising module utilized 3 stacked TF-GridNet blocks. Within each TF-GridNet block, the kernel sizes and strides for all Conv2D and Deconv2D layers were set to $(3, 3)$ and $(1, 1)$, respectively. The embedding dimension for each TF bin was set to 32, the number of hidden units in the BLSTM layers was 128, and the self-attention mechanism was configured with 4 heads and the unfolding layer employed a kernel size of 4 and stride size of 1. For Stage 2, the TF-Restormer configuration followed the original implementation in [9], with the exception of the depth configuration: we utilized 1 encoder layer and 4 decoder layers. Training was performed with a batch size of 4 for Stage 1 and 2 for Stage 2, using 6-second audio segments. AdamW was used for optimization, with an initial learning rate of 1×10^{-3} for Stage 1 (trained for 40 epochs) and 5×10^{-4} for Stage 2 (trained for 10 epochs). Both stages employed cosine annealing learning-rate scheduling, where the rate decays to 0.1 of its initial value.

Table 1. Ablation results on the RASE 2026 validation set.

Method	PESQ	ESTOI	DNSMOS	CS-MFCC
Baseline	1.198	0.153	1.279	0.551
Stage 1 w/o-align.	1.273	0.196	1.963	0.576
Stage 1	1.356	0.214	2.558	0.576
Stage 1 + Stage 2	1.442	0.247	2.607	0.578

Table 2. Final results on the RASE 2026 blind test set. Note that (a) is for direct diaphragm vibration, where the radar measures vibrations from the loudspeaker diaphragm, and (b) is for secondary-surface vibration, where a thin aluminium foil placed near the loudspeaker vibrates indirectly, generating a weaker signal.

Rank	Team	PESQ	ESTOI	DNSMOS	CS-MFCC	Avg. Score (a)	Avg. Score (b)	Overall Score
1st	WHU_IASP	1.541	0.248	2.650	0.598	0.4129	0.2966	0.3431
-	Baseline	1.279	0.166	1.436	0.604	0.2820	0.1994	0.2324

3.2. Results

Table 1 presents ablation results on the RASE 2026 validation set, comparing Stage 1 model without alignment (w/o-align.), Stage 1 with alignment, and the full two-stage RADARNet. Incorporating the alignment strategy in Stage 1 yielded a clear performance gain, confirming that handling the radar–microphone misalignment led to more accurate baseband reconstruction. Furthermore, Stage 2 provides additional gains across all metrics by enhancing the high-frequency spectral content. Table 2 reports the final results on the RASE 2026 blind test set. Our full model, RADARNet, consistently outperformed the baseline[1] in PESQ, ESTOI, and DNSMOS. These test results demonstrated that the proposed two-stage architecture, combined with the alignment strategy, generalized well to test data and effectively improved both speech quality and intelligibility.

4. CONCLUSION

In this paper, we propose RADARNet, a cascaded denoising alignment and restoration framework, to address the ICASSP 2026 RASE Challenge. Our approach explicitly resolves radar–microphone misalignment during training through an alignment-aware denoising front-end. This is followed by a second stage that reconstructs high-frequency details from low-band spectral information. This approach enables RADARNet to jointly and effectively recover fundamental spectral structures and restore missing high-frequency components from noisy radar signals, ultimately achieving superior reconstruction performance.

[1]https://github.com/RASE-Challenge/challenge_baseline2026

5. REFERENCES

[1] A. W. H. Khong, P. A. Naylor, Z. W. Tan, V. G. Reju, R. C. Tewari, and R. Ding, "ICASSP 2026 Radar Acoustic Speech Enhancement Challenge," 2026.

[2] X. Yang, H. Meng, X. Qu, and W. Gao, "Through-the-wall radar imaging grating-lobe and sidelobe suppression method based on imaginary sign coherence factor," *IEEE SPL*, vol. 31, pp. 2120-2124, 2024.

[3] M. Z. Ozturk, C. Wu, B. Wang, and K. R. Liu, "Sound recovery from radio signals," in *Proc. ICASSP*, 2021.

[4] Z. W. Tan, V. G. Reju, R. C. Tewari, R. Ding, and A. W. H. Khong, "Joint enhancement and bandwidth extension for radar through-barrier speech acquisition," *IEEE SPL*, vol. 33, pp. 176-180, 2025.

[5] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," *MobiCom*, 2015.

[6] 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," *MobiSys*, 2019.

[7] M. Z. Ozturk, C. Wu, B. Wang, and K. R. Liu, "RadioMic: Sound sensing via radio signals," *IEEE IoT J.*, vol. 10, no. 5, pp. 4431–4448, 2022.

[8] Z. Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE-ACM TASLP*, vol. 31, pp. 3221–3236, 2023.

[9] U. H. Shin, J. Ko, W. Jeong, and H.-M. Park, "TF-Restormer: complex spectral prediction for speech restoration," *arXiv preprint arXiv:2509.21003*, 2025.

[10] Z. Hou, T. Lei, Q. Hu, Z. Cao, and J. Lu, "SNR-progressive model with harmonic compensation for low-SNR speech enhancement," *IEEE SPL*, vol. 32, pp. 476–480, 2024.

[11] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE-ACM TASLP*, vol. 25, no. 3, pp. 674–686, 2017.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE TASSP*, pp. 320–327, 1976.

[13] A. Gupta, and J. Wu, J. Deng and F. F. Li, "Siamese masked autoencoders," in *Proc. NeurIPS*, vol. 36, pp. 40676–40693, 2023.