Using Cross-modal Distillation to Improve mmWave-based Speech Recognition

Yinnan Zhou*, Lingyu Wang*, Hao Zhou*, Qiyue Li†, Jie Li§, Yusheng Ji‡
*LINKE Lab, School of Computer Science and Technology, University of Science and Technology of China
†School of Electrical Engineering and Automation, Hefei University of Technology
§School of Computer Science and Information Engineering, Hefei University of Technology
‡Information Systems Architecture Science Research Division, National Institute of Informatics
‡Informatics Program, The Graduate University for Advanced Studies, SOKENDAI
Email: {yinnanzhou, lywang19}@mail.ustc.edu.cn, kitewind@ustc.edu.cn,
liqiyue@mail.ustc.edu.cn, lijie@hfut.edu.cn, kei@nii.ac.jp

***Abstract*—As Society 5.0 advances toward human-centric intelligent services, speech interfaces are increasingly deployed in real-world systems, yet traditional microphone-based speech recognition is vulnerable to acoustic noise and suffers from privacy issues. mmWave-based speech recognition is regarded as a promising alternative, but its performance is constrained by the scarcity of large-scale mmWave speech data and the limited information available in the mmWave modality. To address this challenge, we propose mmHear, a mmWave-based system for speech recognition. mmHear exploits the correlation between microphone speech and mmWave speech measurements in the spectrogram domain, and transfers audio-domain knowledge to strengthen mmWave learning under limited mmWave data. Specifically, mmHear pre-trains an audio teacher on large-scale public speech data, adapts it with a small amount of in-domain microphone recordings, and distills its knowledge into a lightweight mmWave model using synchronized microphone–mmWave pairs. We implement a prototype system and evaluate it in realistic environments. Experimental results show that mmHear achieves an average recognition accuracy of 95.6%.**

***Index Terms*—mmWave Sensing, Speech Recognition, Knowledge Distillation**

I. Introduction

With the vision of Society 5.0, cyber–physical integration is expected to deliver human-centric intelligence in everyday settings while preserving information security and privacy. Speech interfaces are a representative example. With the rapid proliferation of voice assistants and speech-driven human–computer interaction (e.g., smart speakers, in-vehicle assistants, and wearable devices), speech understanding has become a key building block in ubiquitous computing systems [1]. It supports a broad range of applications, including smart-home control [2], in-vehicle navigation and infotainment [3], and real-time meeting transcription [4]. Speech interfaces are also increasingly adopted in healthcare [5] and eldercare [6] to enable convenient, low-effort interaction for patients and caregivers. Despite this importance, mainstream solutions remain largely microphone-centric. While microphones are convenient and can be accurate in clean conditions, performance often degrades under ambient noise and interfering sound sources [7]. More critically for security- and privacy-sensitive Society 5.0 deployments, microphones capture airborne acoustics directly, making them vulnerable to replay attacks [8] and raising persistent privacy concerns in always-on scenarios [9].

Fig. 1. mmWave-Based Speech Recognition.

To overcome these limitations, recent research has explored speech-related sensing beyond microphones. Prior work has investigated alternative modalities such as vision [10], WiFi [11], and ultrasound [12] to capture speech-related cues under challenging acoustic conditions, yet these modalities face practical constraints such as privacy implications, limited resolution for capturing fine-grained vibration, or restricted sensing range. Among these alternatives, millimeter-wave (mmWave) is particularly attractive because it is capable of capturing subtle motion and vibration signatures [13], [14]. As shown in Fig. 1, in a typical speech recognition scenario, the mmWave radar captures speech-induced throat vibration signals and recognizes the spoken content via a neural network, with immunity to interference from a nearby loudspeaker.

Existing mmWave-based systems have demonstrated the feasibility of speech-related inference in realistic environ-

The research is partially supported by China National Natural Science Foundation under Grant 62172379 and 625B2174, and partially by JST ASPIRE (JPMJAP2325) and JSPS KAKENHI Grant No. JP24K02937. The authors would like to thank Information Science Laboratory Center of USTC for the hardware services. Hao Zhou is the corresponding author.

ments [15]. However, achieving consistently high recognition accuracy with mmWave remains challenging in practice. A major practical barrier lies in the difficulty of acquiring large-scale, high-quality mmWave training data: unlike microphone speech, which benefits from abundant open-source corpora and mature data-collection pipelines, mmWave speech recording typically requires specialized hardware, careful positioning, and synchronized sensing, making it costly to scale [16]. Consequently, when a mmWave model is trained purely within the mmWave modality using a small dataset, it may learn insufficient and biased latent features and exhibit limited generalization, especially for fine-grained speech recognition.

In this paper, we propose mmHear, a mmWave-based system that leverages audio-domain learning to enhance mmWave speech recognition. mmHear is motivated by a simple observation: microphone speech and mmWave speech signals exhibit certain similarity and correlation in their spectrogram representations; therefore, the audio modality—benefiting from rich open-source datasets and information-rich speech signals—can be leveraged to improve mmWave-based speech recognition performance. mmHear adopts a teacher–student learning paradigm to transfer speech-discriminative knowledge from the audio modality to the mmWave modality using a limited amount of synchronized microphone–mmWave recordings [17]. Specifically, we first pre-train an audio teacher model on large-scale audio data and then adapt it to the target recognition task using a small amount of in-domain microphone recordings. We subsequently distill the teacher's knowledge into a lightweight mmWave model using synchronized audio–mmWave pairs, enabling the mmWave model to learn more robust and discriminative representations than training purely within the mmWave modality.

Our major contributions are summarized as follows.

1) We present mmHear, a mmWave-based system for speech sensing that leverages phase-based mmWave measurements and uses synchronized microphone signals as training-time supervision to improve mmWave speech recognition.
2) We develop a practical teacher–student distillation framework that transfers speech-discriminative knowledge from the audio modality to the mmWave modality, effectively mitigating the data scarcity challenge for mmWave.
3) We build a prototype system and conduct experiments in realistic environments. mmHear achieves 95.6% average recognition accuracy and delivers highly competitive performance on mmWave-based speech recognition.

II. Related Work

A. Microphone-based Speech Recognition

Microphone-based speech recognition has become the dominant solution for voice-enabled applications due to the ease of audio acquisition and the maturity of models and datasets [1]. However, recognition accuracy can degrade notably under ambient noise and competing speakers [7]. In addition, microphone-based systems are susceptible to malicious acoustic commands [18] and raise persistent privacy concerns in always-on scenarios [9].

B. Non-microphone Speech Recognition

To address the limitations of microphones, prior work has explored speech inference using non-audio signals. Chung et al. [10] showed that lip movements can support speech understanding via deep learning. Wang et al. [11] demonstrated that WiFi reflections contain fine-grained motion information for speech inference. Kimura et al. [12] presented an ultrasound imaging approach for silent speech interaction. These approaches, however, face modality-specific constraints: camera-based solutions may raise privacy concerns; WiFi signals provide limited spatial resolution, making fine-grained throat vibration capture difficult; and ultrasound-based approaches often require very close sensing distances.

mmWave-based methods are appealing due to their sensitivity to subtle motion and vibration signatures. Fan et al. [15] proposed mmMIC, which extracts speech-related cues from a single mmWave channel by separately capturing lip-region motion and vocal-cord vibration, and then fuses them for recognition, achieving 92.8% average recognition accuracy. Despite these advances, mmWave-based speech inference still faces practical challenges due to high data collection cost and the scarcity of large-scale open-source mmWave speech datasets, leaving substantial room for improvement [16].

III. Preliminaries

A. FMCW Sensing and Phase-Based Vibration

mmWave radars transmit FMCW chirps and use the phase of reflected signals to sense tiny vibrations. The transmitted signal can be written as

$$s_{\text{tx}}(t) = \exp\left(j2\pi\left(f_c t + \frac{B}{2T}t^2\right)\right), \tag{1}$$

where f_c is the carrier frequency, B is the chirp bandwidth, and T is the chirp duration.

After reflection from a target at distance R, the received signal is a delayed version of the transmitted signal:

$$s_{\text{rx}}(t) = \alpha \exp\left(j2\pi\left(f_c(t-\tau) + \frac{B}{2T}(t-\tau)^2\right)\right), \tag{2}$$

where $\tau = \frac{2R}{c}$ is the round-trip propagation delay, c is the speed of light, and α denotes the attenuation factor. By mixing the transmitted and received signals, an intermediate-frequency (IF) signal is obtained.

After applying a range FFT, reflections from different distances are separated into range bins. We select the bin corresponding to the target region by locating the peak and extracting the complex-valued signal from that bin across consecutive chirps, forming a slow-time complex signal.

When a speaker is located in front of the radar, speech-related micro-motions like throat vibration cause a small, time-varying displacement $d(t)$ along the radar's radial direction. This displacement induces a corresponding phase variation in the reflected signal, which can be approximated as

$$\phi(t) = \frac{4\pi}{\lambda} d(t), \tag{3}$$

where λ is the mmWave wavelength. Owing to the short wavelength of mmWave signals, even sub-millimeter vocal-cord vibrations result in measurable phase changes. Therefore, speech-induced vibration patterns can be captured through phase variations of the mmWave signal, which correlate with the underlying speech activity.

B. mmWave Vibration Spectrum

Let $s(t)$ denote the complex-valued slow-time signal extracted from the selected range bin. We compute its instantaneous phase as

$$\phi(t) = \angle s(t). \tag{4}$$

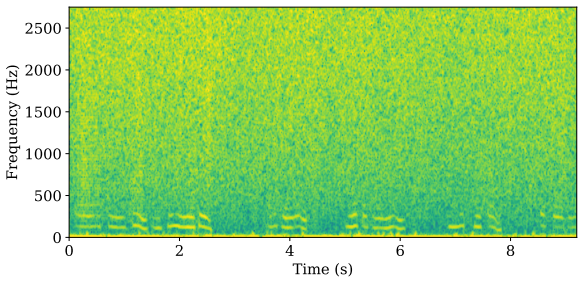
To obtain a stable phase trajectory, we apply standard phase unwrapping, followed by a phase-change operation (e.g., first-order difference) to emphasize speech-induced variations:

$$\Delta\phi(t) = \phi(t) - \phi(t-1). \tag{5}$$

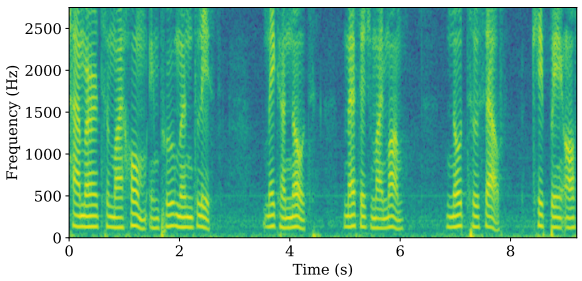
Speech-induced micro-vibrations typically appear as higher-frequency, smaller-amplitude components. Such characteristics are more salient in the resulting phase-change signal and its time–frequency representation. We apply a lightweight temporal filter on $\Delta\phi(t)$ to retain vibration-related components and compute a short-time Fourier transform (STFT) on the filtered signal to obtain

$$X_{\text{mm}}(f,\tau) = \text{STFT}(\Delta\phi(t)). \tag{6}$$

This produces a time–frequency spectrum that characterizes the vibration-induced dynamics of the phase-change signal. We use $X_{\text{mm}}(f,\tau)$ as the mmWave-side input representation to the neural network. Fig. 2a shows an example time–frequency spectrum after processing the vibration signal captured by mmWave.



(a) mmWave vibration spectrogram.



(b) Microphone audio spectrogram.

Fig. 2. Paired spectrograms for the same utterance.

C. Microphone Audio Spectrum

For synchronized microphone audio, we apply the same STFT configuration to obtain a time–frequency spectrum and use the audio-side spectrum as the input representation. Fig. 2b shows an example spectrogram after processing the speech signal captured by the microphone. The paired spectrograms correspond to the same utterance in different modalities and exhibit clear correlation, which provides a foundation for the subsequent cross-modal distillation learning.

IV. System Design

A. Overview

mmHear leverages audio-domain learning to strengthen mmWave-based speech recognition via cross-modal teacher–student training, as illustrated in Fig. 3. In our system, microphone and mmWave signals are first processed into time–frequency spectrum, and mmHear adopts a three-stage training design that progressively transfers speech-discriminative knowledge from the audio modality to the mmWave modality. First, mmHear pre-trains an audio teacher on a large public speech dataset to learn general speech representations from data-rich audio supervision. It then adapts the teacher to the target label space using in-domain microphone recordings, producing a reliable supervisor under our task setting. Finally, mmHear trains a lightweight mmWave model using synchronized microphone–mmWave pairs, where the adapted teacher is fixed and guides the student through a joint objective that integrates three complementary losses for effective cross-modal knowledge transfer.

B. Model

Input representation. mmHear processes microphone and millimeter-wave (mmWave) signals into spectrograms, used as inputs of the following neural network. The detailed signal processing procedures are described in Section III.

Teacher–student architecture. mmHear adopts a classic convolutional backbone for both teacher and student. We use ResNet-18 as the audio teacher, as it is a widely used and well-validated architecture that can learn strong speech-discriminative representations with stable training behavior. To enable efficient mmWave inference and practical deployment, we use a lighter ResNet-10 as the mmWave student, which reduces model complexity while retaining sufficient capacity for the target recognition task. Both networks end with a task-specific classification head, and we use the penultimate-layer feature vector as the latent embedding for representation-level distillation in Stage III.

C. Multi-Stage Training

mmHear trains the teacher and student models in three stages.

Stage I: Audio teacher pre-training on public speech data. mmHear first constructs an audio teacher that can provide reliable supervision for cross-modal transfer. We adopt ResNet-18 as the teacher backbone and pre-train it on the classic dataset (TIMIT) using supervised phoneme classification.

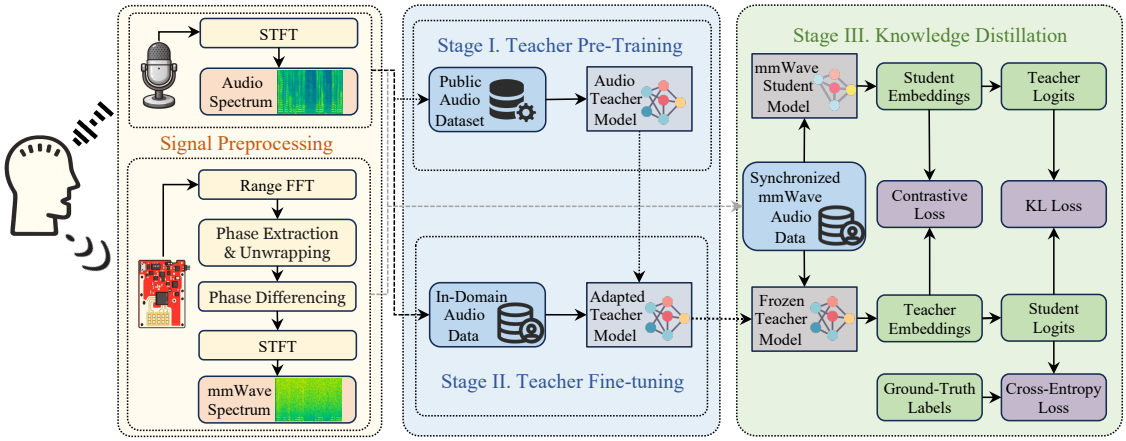


Fig. 3. System Overview.

Each utterance segment is converted into a time–frequency spectrum and the teacher is optimized with the standard cross-entropy loss. This stage equips the teacher with general speech-discriminative representations learned from large-scale public audio data, which serve as a strong initialization for subsequent adaptation.

Stage II: Teacher adaptation to the target label space. The pre-trained teacher is not directly applicable to our target task because TIMIT uses a different label set. We therefore adapt the teacher to our 48-class IPA recognition task using the microphone modality of our synchronized dataset. Concretely, we replace the final classification layer with a new 48-way head and fine-tune the network on our in-domain microphone phase spectrograms.

To make adaptation stable and data-efficient, we use a progressive fine-tuning strategy. We first freeze the feature extractor and train only the new 48-way classification head for a small number of epochs, which quickly aligns the output space to the IPA-48 labels without disrupting the pre-trained representation. We then unfreeze the last two residual stages and continue fine-tuning with a smaller learning rate, allowing the high-level features to adapt to the target recording conditions and the IPA-specific class boundaries while preserving general acoustic features in early layers. This progressive unfreezing schedule yields a strong and well-calibrated audio teacher that is subsequently fixed and used to supervise the mmWave student in the next part.

Stage III: mmWave student training via cross-modal distillation on paired data. After Stage II, mmHear obtains an accurate audio teacher that is aligned with the IPA speech recognition task. Stage III trains a mmWave student to perform the same task using mmWave inputs only, while leveraging the synchronized microphone stream as training-time supervision. For each training sample, we have a synchronized triple (x_a, x_m, y), where x_a is the audio phase spectrogram, x_m is the mmWave phase spectrogram, and $y \in \mathcal{Y}$ is the ground-truth IPA label. The teacher is kept frozen. Given x_a, it outputs teacher logits z_t and a penultimate-layer feature embedding h_t. Given x_m, the student outputs student logits z_s and an embedding h_s. The student is trained with three complementary losses.

Cross-entropy loss. The student is directly optimized to predict the ground-truth IPA label. Let p_s denote the student posterior distribution obtained from z_s. Using the one-hot label vector y, the supervised loss is

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c\in\mathcal{Y}} y_c \log p_{s,c}. \tag{7}$$

KL divergence loss. To transfer the teacher's soft decision structure, mmHear distills the teacher's posterior distribution into the student. We apply temperature scaling with τ to obtain softened distributions [17]

$$p_t^\tau = \mathrm{softmax}(z_t/\tau), \qquad p_s^\tau = \mathrm{softmax}(z_s/\tau), \tag{8}$$

and define the distillation loss as

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(p_t^\tau \,\|\, p_s^\tau) = \sum_{c\in\mathcal{Y}} p_{t,c}^\tau \log \frac{p_{t,c}^\tau}{p_{s,c}^\tau}. \tag{9}$$

This term encourages the mmWave student to follow the teacher's class posterior distribution on synchronized samples, providing richer supervision than hard labels alone.

Contrastive loss. Logit-level distillation aligns predictions, but cross-modal transfer can further benefit from aligning latent representations. We therefore apply a cross-modal InfoNCE objective on the penultimate embeddings. For a mini-batch of size N, denote the teacher and student embeddings as $\{h_t^i\}_{i=1}^N$ and $\{h_s^i\}_{i=1}^N$, and use ℓ_2-normalized features so that the dot product corresponds to cosine similarity. For each anchor i, we define the in-batch positive set

$$\mathcal{P}(i) = \{j \mid y_j = y_i\}. \tag{10}$$

The contrastive loss is then formulated as

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{|\mathcal{P}(i)|} \sum_{j\in\mathcal{P}(i)} \log \frac{\exp\left(h_s^i \cdot h_t^j/\gamma\right)}{\sum_{k=1}^{N} \exp\left(h_s^i \cdot h_t^k/\gamma\right)}, \tag{11}$$

where γ is the contrastive temperature. This objective encourages embeddings of the same IPA label to be close across modalities and separates embeddings from different labels.

The student model is trained by jointly optimizing a weighted combination of the three complementary losses, formulated as $\mathcal{L}_{\text{total}} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$. By integrating these objectives into a unified training criterion, the cross-modal distillation pipeline enables effective knowledge transfer from the adapted audio teacher to the mmWave student. It stabilizes optimization and promotes alignment in both prediction space and latent feature space, leading to more discriminative and robust representations for mmWave-based speech recognition.

V. Experimental Evaluation

A. Experiment Setup

Implementation. As shown in Fig. 4, our prototype is built on an IWR6843ISK mmWave radar, a DCA1000EVM data capture board, a microphone, and a laptop. The radar operates as a 60 GHz FMCW sensor. During the data collection process, the equipment is positioned 50 cm from the participant. In our configuration, the chirp period is 100 μs, and each chirp contains 80 ADC samples. The training and inference processes are conducted on a server with an AMD EPYC 7763 CPU and an NVIDIA GeForce RTX 4090 GPU.

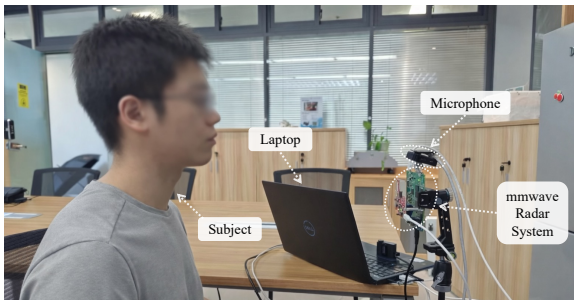


Fig. 4. Experiment setup.

Dataset. We use both a public audio corpus and a self-collected synchronized dataset. For teacher pre-training, we use the TIMIT dataset and follow its standard phoneme recognition setup with 61 phoneme classes. For fine-tuning and cross-modal learning, we collect a synchronized microphone–mmWave dataset using the classic 48 phonetic symbols as recognition objects. Each sample corresponds to an isolated utterance segment of one target symbol. The dataset contains recordings from 10 volunteers (7 males and 3 females) and comprises 9,600 paired samples, including microphone speech signals and the corresponding mmWave vocal-cord vibration measurements. We split this dataset into 80% training and 20% testing, and ensure that both splits include all target symbol classes.

Metrics. We evaluate recognition performance using accuracy and recall. Accuracy is defined as the fraction of correctly classified samples among all test samples. Recall is computed per class as the fraction of correctly classified samples over the total samples of that class, and we report the macro-averaged recall by averaging recall values equally across all classes.

B. Performance

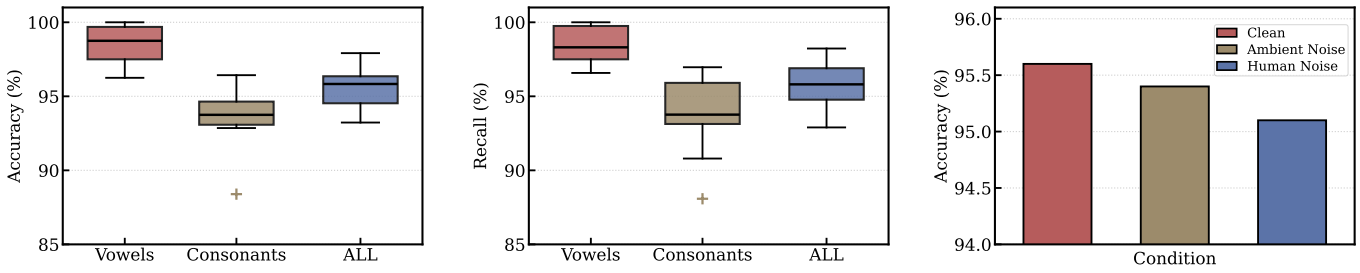
1) Overall Performance: mmHear demonstrates strong speech recognition performance in our evaluation. Fig. 5 presents subject-wise boxplots across 10 subjects, summarizing recognition results on vowels, consonants, and all phonetic symbols. As shown in Fig. 5a, mmHear achieves an average accuracy of 95.6% on the test set, indicating that cross-modal training effectively strengthens mmWave-based speech recognition. Vowel recognition achieves about 98% accuracy on average, while consonant recognition is comparatively more challenging, with the lowest-performing subject reaching around 88%. Fig. 5b further reports an average macro recall of 95.7%.

2) Performance Across Training Stages: mmHear is trained in three stages, and we report the performance after each stage to illustrate the progressive transfer of audio-domain knowledge to the mmWave modality. After supervised pre-training on TIMIT with 61 phoneme classes, the audio teacher achieves 82.4% accuracy on the TIMIT test split. Although this accuracy may not appear high, the reason is that TIMIT phoneme labels are derived from segments cut out of continuous sentences rather than isolated phoneme recordings, which makes the classification setting inherently less clean than isolated phoneme recognition. Therefore, the teacher can still acquire useful speech-discriminative representations from large-scale public audio data in this stage. We then replace the classification head and fine-tune the teacher on our in-domain microphone recordings for the target phonetic symbol recognition task; the adapted teacher reaches 98.7% accuracy on the microphone test set and provides reliable supervision for cross-modal transfer. Finally, we distill the teacher into the student using paired microphone–mmWave samples, and the mmWave student model attains 95.6% recognition accuracy. These results show that the staged training procedure effectively leverages public audio data and in-domain adaptation to strengthen mmWave recognition.

3) Impact of Noise: We evaluate the robustness of the system under two common noise conditions. For ambient noise, we play background music at approximately 80 dB near the subject during data collection. For human noise, we introduce interference from nearby people who walk around the subject and speak loudly in the vicinity. As shown in Fig. 5c, mmHear maintains stable recognition performance in both settings, exhibiting no noticeable degradation compared to the clean condition. These results demonstrate the excellent noise resistance of mmHear.

C. Ablation Study

We conduct ablation experiments to examine how audio-domain supervision contributes to mmHear, focusing on three key aspects: the role of teacher guidance, the impact of large-scale audio pre-training, and the effect of the distillation loss design.



(a) Accuracy (b) Recall (c) Impact of Noise

Fig. 5. Performance.

Impact of teacher guidance. Impact of teacher guidance. We remove the audio teacher and train the mmWave student using only supervised learning on mmWave data, resulting in an accuracy drop to 89.1%. This result shows that the audio-domain guidance is critical for effective mmWave inference.

Impact of teacher pre-training on public audio data. Training the ResNet-18 teacher from scratch using only in-domain microphone recordings, without TIMIT pre-training, reducing the student accuracy to 93.4%. This degradation stems from the lack of exposure to broader audio variations, which increases overfitting to the limited in-domain distribution and weakens cross-modal knowledge transfer.

Impact of loss design in cross-modal distillation. We further analyze the contribution of different loss components. Using cross-entropy and KL divergence yields 90.5% accuracy, while using cross-entropy and contrastive loss result in 93.8%. This result suggests that representation-level alignment provides stronger supervision than logit-level distillation in mmHear. These two losses are complementary, and the best system performance is achieved when all three losses are jointly optimized.

VI. Conclusion

In this paper, we presented mmHear, a mmWave-based system for speech recognition. mmHear builds a teacher–student learning framework that uses synchronized microphone recordings as training-time supervision and transfers speech-discriminative knowledge from the audio modality to the mmWave modality. We implemented a prototype system and evaluated mmHear in realistic environments, achieving an average recognition accuracy of 95.6%. The experimental results demonstrated the effectiveness of cross-modal knowledge transfer for enhancing mmWave-based speech recognition.

References

[1] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] C. Geeng and F. Roesner, "Who's in control? interactions in multi-user smart homes," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–13.

[3] Y. Zheng, Y. Liu, and J. H. L. Hansen, "Navigation-orientated natural spoken language understanding for intelligent vehicle dialogue," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 559–564.

[4] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang *et al.*, "Advances in online audio-visual meeting transcription," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 276–283.

[5] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.

[6] P. Upadhyay, S. Heung, S. Azenkot, and R. N. Brewer, "Studying exploration & long-term use of voice assistants by older adults," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–11.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[8] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the 4th ACM workshop on security and privacy in smartphones & mobile devices*, 2014, pp. 63–74.

[9] S. Gray, "Always on: privacy implications of microphone-enabled devices," in *Future of privacy forum*, 2016, pp. 1–10.

[10] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[11] 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*, 2014, pp. 593–604.

[12] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–11.

[13] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[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*, 2019, pp. 14–26.

[15] L. Fan, L. Xie, X. Lu, Y. Li, C. Wang, and S. Lu, "mmmic: Multi-modal speech recognition based on mmwave radar," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

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

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[18] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX security symposium (USENIX security 16)*, 2016, pp. 513–530.