FeelWave: Enabling Emotion-Aware Voice Interaction through Noise-Robust mmWave Emotion Sensing

Lingyu Wang
University of Science and Technology of China
Hefei, Anhui, China
lywang19@mail.ustc.edu.cn

You Zuo
University of Science and Technology of China
Hefei, Anhui, China
leftright@mail.ustc.edu.cn

Dequan Wang
University of Science and Technology of China
HeFei, Anhui, China
wdq15588@mail.ustc.edu.cn

Chenming He
University of Science and Technology of China
Hefei, Anhui, China
hechenming@mail.ustc.edu.cn

Chengzhen Meng
University of Science and Technology of China
Hefei, Anhui, China
czmeng@mail.ustc.edu.cn

Xinran Zhang
University of Science and Technology of China
Hefei, Anhui, China
zxrr@mail.ustc.edu.cn

Xiaoran Fan
Independent Researcher
Sunnyvale, California, USA
vanxf@google.com

Yanyong Zhang
University of Science and Technology of China
Hefei, Anhui, China
yanyongz@ustc.edu.cn

Abstract

Voice has been a primary interaction mode with LLM-powered assistants. Beyond semantics, voice carries emotional cues with potential to guide empathetic system responses. Yet, robust vocal emotion sensing in noise and its use in optimizing interactions remain underexplored. In response, we present *FeelWave*, which achieves empathetic voice interaction through noise-robust mmWave emotion sensing and structured LLM prompts. It extracts robust vocal information from mmWave signals, applies audio-to-mmWave transfer learning for efficient emotion recognition, and employs chain-of-thought-based query optimization to enable emotion-adaptive responses. Evaluations show that *FeelWave* achieves 92.3% emotion recognition accuracy and remains robust in noisy environments, yielding a 62.9 percentage-point gain over audio-based models. In voice interaction studies, 74.3% of users prefer *FeelWave*, reporting significantly higher satisfaction than a baseline without emotion sensing (4.37 vs. 3.22). A SUS score of 88.3 confirms *FeelWave*'s high usability in real-world deployment. We hope this work will inspire more empathetic, user-centered AI-driven assistants.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Human computer interaction (HCI)**.

Keywords

Emotion, Voice Interaction, mmWave

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790631

ACM Reference Format:
Lingyu Wang, You Zuo, Dequan Wang, Chenming He, Chengzhen Meng, Xinran Zhang, Xiaoran Fan, and Yanyong Zhang. 2026. FeelWave: Enabling Emotion-Aware Voice Interaction through Noise-Robust mmWave Emotion Sensing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3790631

1 INTRODUCTION

Voice, as a natural carrier of user intent, has been a primary mode of interaction with intelligent agents powered by large language models (LLMs), such as Li Auto's MindGPT [9] and Apple Intelligence [8]. Prior studies [60, 110] show that emotional cues in texts can enhance LLM generation, indicating similar potential in voice to enable empathetic responses. However, current LLMs largely confine emotion analysis to semantics [98], whereas real-world voice interactions typically involve emotionally neutral commands [101, 106]. Beyond semantics, voice universally conveys emotions through tone, volume, and rhythm, making it pervasive across diverse conversations [59]. Imagine an emotion-aware agent that not only understands the query but also adapts to the user's emotion. As shown in Fig. 1, the agent provides clear and reassuring guidance when a user anxiously asks about an interview despite the neutral semantics, and responds concisely in a calming manner when an angry driver requests navigation. Our preliminary user study further reinforces this observation: among 25 participants, 84.7% prefer emotion-integrated LLM interactions, rating them as more contextually relevant, emotionally appropriate, and comfortable than baseline responses without emotional adaptation. *These findings motivate the design of a transparent, emotion-adaptive voice interaction system.*

To this end, this paper explores the feasibility of transparently leveraging user emotions from vocal features to design an emotion-adaptive voice interaction system that enhances user experience. As

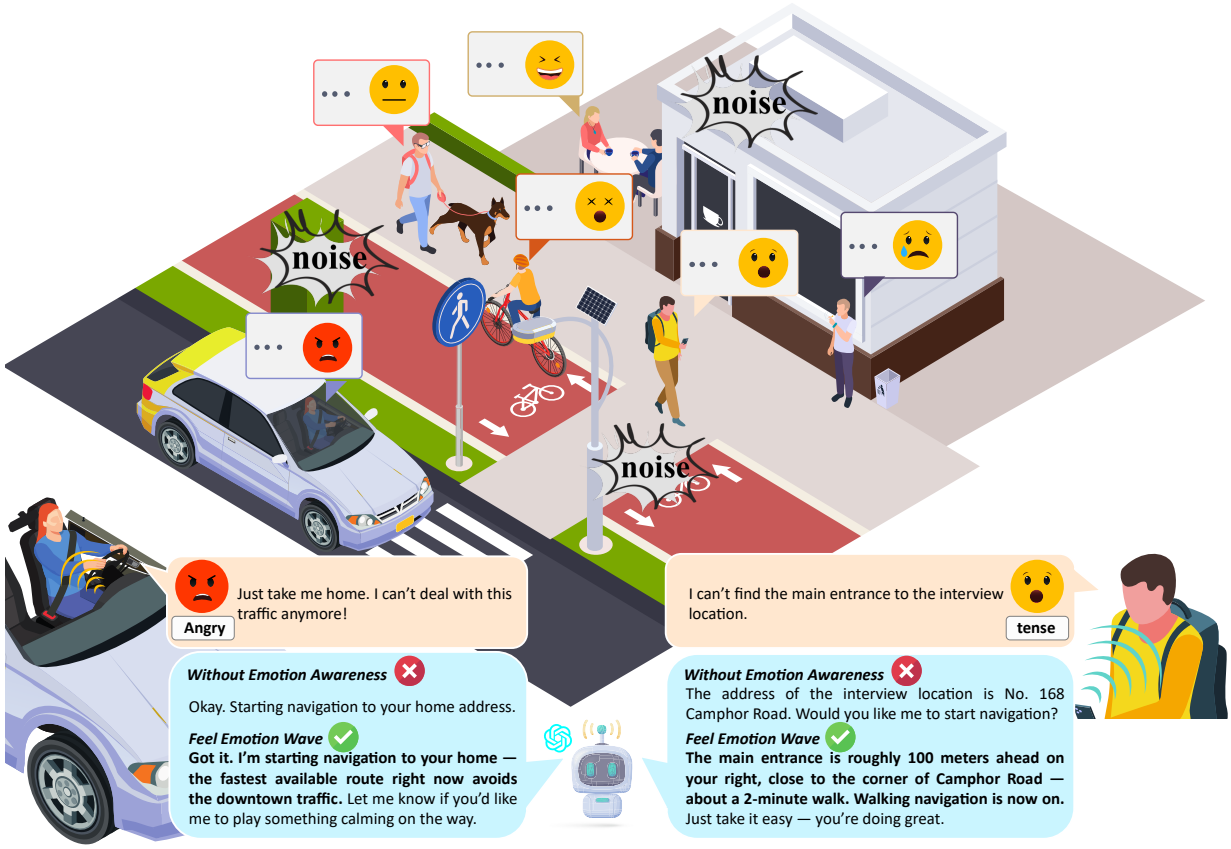

Figure 1: Applications of *FeelWave* in daily life. Users can interact with smart devices equipped with *FeelWave* (e.g., smartphones, watches, and cockpits) for emotion-aware voice interaction.

illustrated in Fig. 1, the system enables personalized and emotionally responsive interaction in smart cockpits and mobile devices. To achieve this vision, the system is guided by three design goals. ***(1) Robust emotion sensing.*** It reliably captures vocal features to enable accurate emotion recognition under diverse conditions, including background noise and natural body movement. ***(2) Lightweight design.*** It supports accurate, real-time emotion recognition with a compact model suitable for deployment on resource-constrained mobile devices. ***(3) Effective integration with LLMs.*** It enables LLM agents to perform emotion-adaptive reasoning, generating responses that meet user needs and provide emotional value, rather than merely appending emotional states to queries, resulting in superficial or tone-deaf interactions.

Although prior work [48, 107] has leveraged vocal emotions to enhance interaction, its reliance on microphones for emotion sensing makes it highly susceptible to airborne noise and unreliable in low signal-to-noise ratio (SNR) conditions [88, 116]. The scarcity of emotion-labeled speech data collected under noisy conditions makes it difficult to develop models that can reliably perceive emotions in real-world environments [35, 57]. While speech enhancement methods [24, 87] improve signal quality, residual noise and spectral distortion still hinder reliable emotion analysis [97, 126]. Our experiments confirm that audio-based emotion recognition methods remain poor even after enhancement, with only 46.02% accuracy at -5 dB SNR. To unobtrusively capture vocal cues while remaining robust to noise, we use millimeter-wave (mmWave) radar for vocal vibration sensing, preserving fundamental frequencies and partial harmonics for effective emotion analysis, as shown in Fig. 2. Our analysis reveals that mmWave-captured vocal vibrations are strongly correlated with microphone-recorded voice, with a cosine similarity of 0.81 in glottal source features that encode rich emotional cues [118]. However, mmWave sensing is sensitive to motion in everyday use, causing performance degradation. Additionally, emotion recognition heavily depends on dataset scale, and the scarcity of mmWave datasets hampers model training, resulting in suboptimal performance. Furthermore, integrating emotional states into LLMs' reasoning for deeper understanding of user needs, beyond superficial responses, remains a challenge.

In this paper, we propose *FeelWave*, an emotion-aware voice interaction system that combines robust mmWave emotion sensing with structured LLM prompts. We design an innovative motion-robust vocal vibration extraction algorithm with single-input single-output (SISO) mmWave radar for unobtrusive, noise-resilient emotion sensing. To mitigate performance loss from limited mmWave data, we design a cross-modal transfer pipeline that maps large-scale audio features to mmWave features. By integrating emotion information into query optimization prompts, LLM agents such as GPT-4o [78] and Gemini [105] can adapt their behaviors to emotional cues, generating responses that are both contextually relevant and empathetic. We implement *FeelWave* as follows.

First, we aim to extract vocal vibration signals via mmWave while remaining robust to motion. A common approach captures

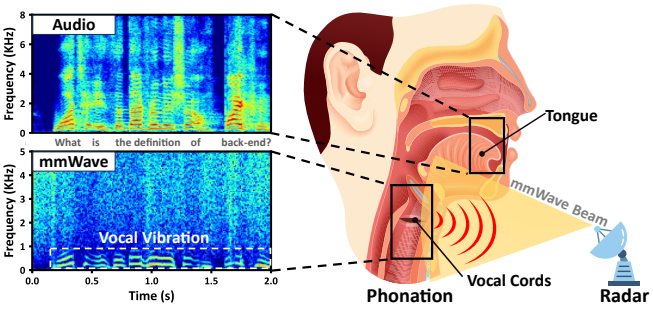

Figure 2: Vocal cord vibrations and the resulting voice during phonation. The mmWave-captured vocal vibrations contain the voice fundamental frequency and partial harmonics.

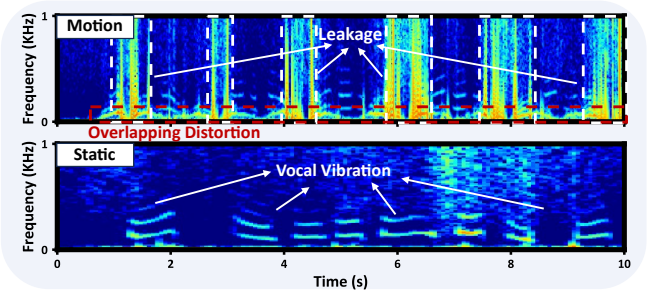

Figure 3: Real-world motion interference. Motion causes signal discontinuities leading to spectral leakage, and modulates the signal, resulting in overlapping distortion.

phase differences at a fixed range bin corresponding to the vocal cords. However, body movement smears vocal signals across multiple range bins, causing dispersed energy, signal discontinuities, and spectral leakage, as shown in Fig. 3. Moreover, motion induces unexpected overlapping within phase differences of each bin, rendering filter-based suppression [27, 87] ineffective. To address this, we propose a novel vocal signal extraction algorithm. We model the statistical distribution of vocal energy to dynamically search within a neighborhood of bins for the vocal-intensive bin over time, rather than relying on a fixed bin. Through an in-depth analysis of intra-bin phase differences, we uncover a previously underexplored mechanism showing that overlapping distortion arises from coupling between body motion and the radar's direct current (DC) component. Leveraging this, we design a motion demodulation strategy that, through adaptive DC removal, corrects such distortions within the selected bin, yielding refined vocal signals.

Moreover, the lack of large-scale mmWave data and the high cost of collection pose challenges to developing an mmWave-based model that is both lightweight and accurate. Fortunately, advances in speech representation learning, enabled by large-scale pretraining on unlabeled audio, support effective transfer to low-resource tasks. Given the strong consistency between mmWave and voice emotion-related source features, knowledge distillation from audio to mmWave offers a feasible solution. However, incomplete harmonics in mmWave vocal signals lead to representation degradation, resulting in suboptimal convergence. To address this, we propose a cross-modal transfer pipeline that acts as a regularizer by distilling knowledge from a speech representation model, where comprehensive audio representations enhance offline convergence quality and mmWave ensures noise-robust online inference. Specifically, we design a hybrid layer-wise distillation loss to mitigate performance degradation caused by the feature deficiency of mmWave.

Furthermore, under different emotional states, can LLMs deeply analyze user needs and provide personalized responses by leveraging their contextual understanding capabilities? Recent work [30, 93] shows that physiological data from wearable devices can enhance the personalization capabilities of LLMs. Inspired by this, we propose a two-stage emotion-driven query optimization module that incorporates emotional information through in-context learning with chain-of-thought (CoT) prompting [63, 112]. Given that insufficient contextual reasoning can lead to superficial responses that overlook users' underlying needs [6, 91], this module enables LLMs to jointly reason over user intent and emotion when optimizing the request, identifying goals such as emotional support, cognitive simplification, task-oriented guidance, and wellness promotion. It then formulates emotion-aligned prompts based on the identified goals, allowing LLMs to generate responses that are both contextually relevant and emotionally adaptive.

To validate the effectiveness of *FeelWave*, we conduct extensive evaluations. On a dataset of 27 participants across six emotions, it achieves 92.28% emotion recognition accuracy, outperforming four state-of-the-art baselines. We further evaluate *FeelWave*'s performance in voice interaction. In a user study with 20 participants, 74.3% prefer voice interactions with *FeelWave*, reporting significantly higher overall satisfaction (4.37 ± 1.23 vs. 3.22 ± 1.03) compared to a baseline without emotion sensing. In in-the-wild evaluations with strong noise and natural movement (e.g., subway, café, driving), *FeelWave* maintains robust performance. A System Usability Scale (SUS) [12] evaluation demonstrates its high usability in the real world, yielding a score of 88.3 ($\geq$ 85 is considered "excellent" [11]). Our dataset, model, and demo video are available at **GitHub**. In conclusion, our contributions are as follows:

- We propose *FeelWave*, an emotion-aware voice interaction system that enables noise-robust emotion sensing and emotion-adaptive LLM response generation. To the best of our knowledge, *FeelWave* is the first to reliably apply mmWave vocal sensing for emotion inference, offering a noise-resilient acoustic alternative to audio and a scalable direction for advancing research on emotion-aware agents. It delivers transparent, empathetic voice interactions that enhance user experience and ensure reliability in real-world environments.
- We design a motion-robust vocal signal extraction algorithm that dynamically captures vocal-intensive signals over time and demodulates motion-induced distortions to produce refined vocal signals, significantly enhancing the reliability of mmWave emotion sensing in dynamic scenarios.
- We develop a cross-modal transfer pipeline that mitigates the intrinsic representational deficiency of mmWave vocal signals via layer-wise feature alignment with audio representations, enabling lightweight yet discriminative emotion inference. We further design a two-stage emotion-driven query optimization module that enhances the LLM's ability to move beyond superficial

reasoning, generating responses that are both contextually relevant and empathetic.

- We validate *FeelWave*'s emotion sensing and voice interaction effectiveness, achieving 92.28% emotion recognition accuracy and robustness in real-world scenarios. Additionally, *FeelWave* is widely preferred in voice interactions (74.3%), with significantly higher overall satisfaction (4.37 ± 1.23) compared to the baseline (3.22 ± 1.03).

2 RELATED WORK

2.1 Emotion Sensing for Voice Interactions

Human emotions are commonly inferred from multimodal cues, including facial expressions [58, 123], vocal tones [26, 35], and physiological signals [121, 124] such as heart rate and respiration. Emotion sensing has therefore become a central research area, with extensive studies devoted to advancing this domain. Picard's seminal work on affective computing [90] envisions computers that can perceive users' emotional states and dynamically adapt their interactions. Yet, developing robust modalities for reliable emotion sensing remains a significant and open challenge.

2.1.1 Challenges of Audio-based Emotion Sensing. Voice has emerged as a primary mode of interaction with smart devices [8, 9], conveying emotional information through variations in tone, volume, and rhythm. In contrast, visual cues such as facial expressions [58, 123] impose higher user burden and often fail to align consistently with semantic content. They are also difficult to capture during subtle emotional shifts and are highly sensitive to lighting, viewing angle, and occlusion (e.g., masks). Physiological signals [2, 56] like ECG require users to remain still, limiting applicability in natural interactions. *In comparison, voice-based emotion sensing is low-burden, natural, and seamlessly integrates with spoken queries, making it ideal for intelligent voice assistants.*

Speech simultaneously carries semantic and emotional information, yet their robustness under noise is asymmetric. Large-scale noisy speech datasets have driven significant progress in noise-robust automatic speech recognition (ASR), enabling models such as GPT-4o-transcribe [83] to achieve reliable performance across diverse conditions in practice [65, 80]. However, the scarcity of emotion-labeled noisy speech data hinders stable emotion perception in real-world environments [35, 57]. Although prior work [48, 107] has leveraged vocal emotions to enhance interaction, its reliance on microphones for emotion recognition makes it highly unreliable in low SNR environments [88, 116]. While speech enhancement methods [24, 87] can enhance signal quality, spectral distortion and residual noise still hinder effective emotion analysis [97, 126]. Recent in-ear sensing studies [39, 40] suppress airborne noise more effectively but rely on contact-based earphones, limiting comfort and suitability for open, natural interactions. Thus, enabling noise-robust emotion sensing through user-transparent vocal features remains an open challenge.

2.1.2 Toward mmWave-based Emotion Sensing. For emotion sensing, prior studies [121, 124] have inferred emotions from physiological cues (e.g., heart rate, respiration) captured by mmWave. However, these cues are easily confounded with activity-induced physiological changes (e.g., walking, climbing stairs) and require users to remain still for multiple cycles [56], limiting their reliability and real-time applicability. Leveraging its immunity to airborne noise and ability to capture fine-grained vocal vibrations [119], mmWave radar shows strong potential for transparent and noise-resilient voice interactions. Prior work has used mmWave vocal sensing for speech reconstruction [119], recognition [27, 66, 125], enhancement [87], and speaker identification [67], demonstrating effective noise robustness. Despite these advances, the tiny wavelength of mmWave (about 4 mm) makes it highly sensitive to user-device motion, leaving motion interference a persistent challenge. Existing approaches above extract vocal vibrations from phase differences within a single range bin, neglecting that large body movements disperse vocal signals across multiple bins, leading to discontinuities and spectral leakage (Fig. 3). Moreover, within each range bin, motion coupling with the radar's DC component induces nonlinear modulation, causing vocal and motion-induced frequencies to indistinguishably overlap, thereby rendering filter-based suppression [27, 87] ineffective. A recent work [14] reduces motion interference in the target region by using reference reflections from other body parts with similar motion patterns, but its reliance on Multi-Input Multi-Output (MIMO) radars for angular resolution increases system cost and latency. Addressing these limitations, we design a novel motion-robust vocal signal extraction algorithm using SISO mmWave radar that dynamically captures vocal-intensive signals from a neighborhood of bins and demodulates motion-induced distortions, yielding refined vocal signals.

Since the high-frequency harmonics in mmWave-captured vocal vibrations are incomplete, it remains uncertain whether mmWave sensing alone can achieve effective emotion recognition. Unlike multimodal fusion with audio or vision, a unimodal mmWave design inherently avoids cross-modal alignment errors and residual noise while greatly reducing computational and deployment costs [3, 119]. This paper is the first to demonstrate the feasibility of noise-resilient emotion recognition from mmWave vocal vibrations.

2.2 LLM-based Personal Agent

The rapid development of LLMs has greatly expanded the capabilities of personal assistants [62]. Recent work [16, 30, 93] integrates LLMs with sensor data to create context-aware agents that enrich user experiences. For instance, WellMax [93] leverages physiological signals from wearable devices to improve personal agent responses. Prior research [60, 110] also shows that emotional prompts can significantly improve the generative performance of LLMs, while Fang et al. [29] demonstrate that emotionally supportive responses can strengthen users' emotional resonance and sense of self-affirmation. Inspired by these insights, *FeelWave* incorporates robust emotion sensing to guide personal agents through emotion-aware prompting, fostering emotional intelligence and enabling more empathetic interactions.

Recent LLMs have advanced in emotional understanding and generation. Psychological assessments [100, 110] show that models such as ChatGPT-4, Gemini 1.5 Flash, and DeepSeek V3 can recognize and reason about complex emotions. However, their emotional intelligence in real-world settings remains limited, as they rely primarily on textual semantics for emotion analysis [98] and struggle

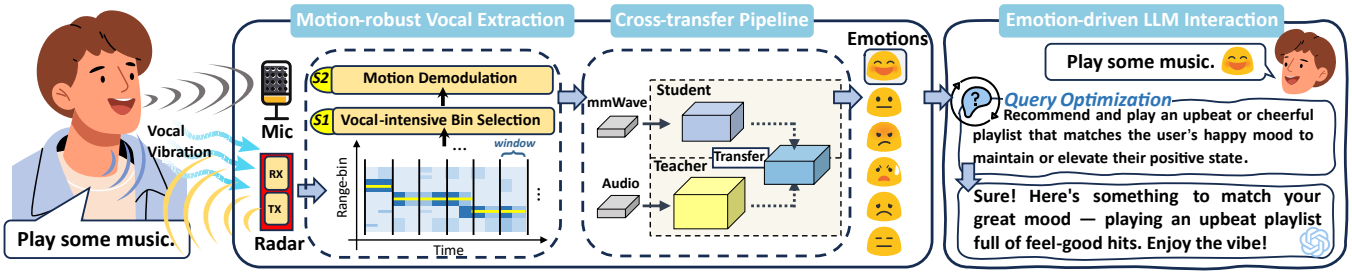

Figure 4: System overview. *FeelWave* consists of three core modules: (1) motion-robust vocal signal extraction, which selects vocal-intensive mmWave signals and compensates for motion to recover clean signals; (2) a cross-transfer pipeline distilling knowledge from an audio teacher into an mmWave student for accurate, lightweight emotion recognition; and (3) emotion-driven LLM interaction, where recognized emotions guide query optimization for empathetic responses.

to generalize across dynamic contexts [91]. In everyday voice interactions with agents, users often convey emotions through acoustic cues such as tone, volume, and rhythm, while textual content remains emotionally neutral due to its predominantly functional nature [101, 106]. To address this gap, recent studies [17, 117] translate acoustic features into textual emotion captions using LLMs to improve the understanding of speaker emotions. Yet, while semantics aid emotion inference in human dialogue [117], their emotional relevance in functional agent commands is minimal. Therefore, given that acoustic features pervasively convey emotional cues across diverse conversational contexts [59], leveraging them alone offers a practical, scalable path to enhancing agents' emotional intelligence. Meanwhile, to enable LLM-powered agents to move beyond purely textual semantics in emotion analysis, other studies [18, 64] explore multimodal integration of audio and visual cues to enhance emotional understanding. However, such fusion is resource-intensive and introduces high latency during real-time interactions. Furthermore, both acoustic-to-text description and multimodal fusion rely on fragile modalities, as audio is noise-sensitive and vision is affected by lighting, angle, and occlusion.

To enable emotionally intelligent personal agents that perform reliably in real-world settings, *FeelWave* integrates mmWave radar to robustly sense users' vocal emotions under noise. Our system develops a plug-and-play emotion detector based on acoustic cues that can seamlessly integrate with arbitrary LLM-powered agents. This does not conflict with emerging efforts in acoustic-semantic fusion and emotion captioning. Instead, the mmWave acoustic cues encoded by *FeelWave* are transferable to these explorations, offering a noise-immune alternative to audio and a scalable foundation for advancing research on emotion-aware agents. We further employ techniques such as chain-of-thought reasoning (CoT) [63, 112] and prompt-based task decomposition [70] as a cost-effective, fine-tuning-free strategy to structure LLM reasoning, enabling emotionally expressive yet contextually grounded responses.

3 FEELWAVE DESIGN

mmWave-captured vocal vibrations align closely with microphone-recorded voice in emotion-rich source features such as pitch, intensity, and rhythm, with a cosine similarity of 0.81. Thus, mmWave vocal vibrations can support noise-robust emotion perception, as detailed in Appendix A.1. Building on this, we propose *FeelWave*, an emotion-aware voice interaction system that enables robust mmWave-based emotion sensing and emotion-driven LLM response generation. As shown in Fig. 4, *FeelWave* comprises a microphone and an mmWave radar. Specifically, we use GPT-4o-transcribe [83] for speech-to-text conversion, which has demonstrated state-of-the-art robustness in noisy environments [65, 80], and GPT-4o-mini-TTS [82] to synthesize voice responses. We mainly focus on mmWave sensing to design three key modules that enable ***robust***, ***lightweight***, and ***emotion-adaptive voice interaction***.

- **Motion-robust Vocal Signal Extraction.** This module demodulates motion-induced distortions to produce refined vocal signals, significantly enhancing the reliability of mmWave emotion sensing in dynamic settings.
- **Cross-transfer Pipeline.** This module uses layer-wise distillation with comprehensive audio representations to offline enhance mmWave convergence quality, enabling lightweight and noise-robust online emotion inference.
- **Emotion-driven LLM Interaction.** This module enhances LLMs via a two-stage emotion-driven query optimization to move beyond superficial reasoning, generating responses that are both contextually relevant and empathetic.

3.1 Motion-robust Vocal Signal Extraction

We use a SISO mmWave radar to extract vocal vibrations from phase differences obtained from reflected IF signals. User movement, however, complicates this process: it disperses vocal energy across multiple range bins, and each range bin spans a spatial region and thus retains residual motion interference. To locate the bin with the richest vocal content, we design a vocal energy estimator. We further propose a motion demodulation algorithm to correct distortions within a single bin and extract refined vocal signals.

3.1.1 Vocal-intensive Bin Selection. User motion progressively disperses the mmWave in-phase and quadrature (IQ) signals from the vocal cord region across nearby range bins, dispersing the vocal information and introducing spectral leakage, as shown in Fig. 5(a). To preserve complete vocal features under motion, we search over a neighborhood of adjacent range bins rather than relying on a single bin, with a size accommodating typical body movement during daily use. The IQ signal whose phase difference contains the richest vocal content is extracted as follows.

Neighborhood Construction. To locate reflected signals most likely to contain vocal vibrations, we first apply Range-FFT [96] to

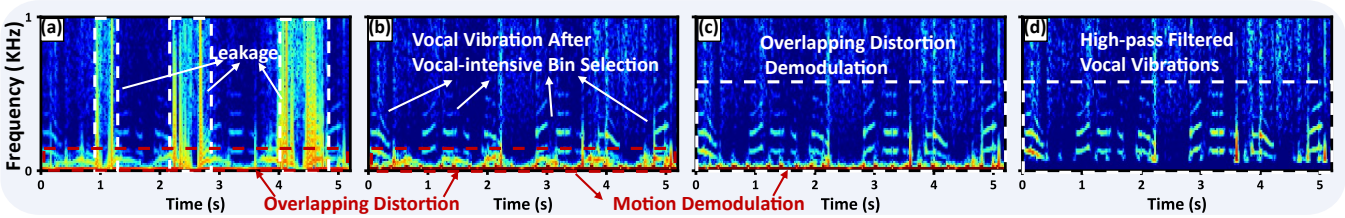

Figure 5: Motion-robust vocal signal extraction. (a) Spectral leakage and overlapping distortion caused by real-world motion interference; (b) Vocal vibrations after vocal-intensive bin selection, with remaining distortion; (c) Vocal vibrations after motion demodulation; (d) Final output after high-pass filtering.

the mmWave IF signals within a time window W, yielding the reflection distribution across range bins. We further limit the analysis to a detection range R to suppress reflections outside the user, forming a range profile as $M \in \mathbb{R}^{R\times W}$. Since the signal with the highest intensity likely carries the most user-related information, we identify the corresponding range bin as $r_{\max} = \text{argmax}_r \sum_{t=1}^{W} |M(r,t)|$, which serves as the anchor for locating vocal-relevant bins. We then locate a neighborhood of size $N = 2n+1$ centered on $r_{\max}$, denoted as $\mathcal{N}_{\text{neigh}} = \{s_1, s_2, \ldots, s_N\}$. This neighborhood provides the candidate region for the subsequent search for the IQ signal richest in vocal content.

Vocal Energy Guided Search. Motivated by the effectiveness of Gaussian-based models in voice activity detection [36], we design a vocal energy estimator that models the statistical distribution of multi-band energy features within $\mathcal{N}_{\text{neigh}}$ to search for the vocal-intensive bin. Since mmWave-captured vocal vibrations lie below 1 kHz, phase differences computed from these bins are respectively decomposed into several frequency sub-bands, and their band-wise energy measures are aggregated into a compact representation. A single Gaussian model characterizes this feature distribution, and the IQ signal with the highest likelihood, denoted as S_{opt}, is selected as containing the most vocal energy. To accommodate gradual changes in vocal characteristics, the model parameters are updated using a slow-learning strategy.

In addition, to eliminate phase discontinuities at the boundaries of consecutive time windows, we apply a smoothing window to S_{opt}, yielding the smoothed IQ signal $\tilde{S}_{\text{opt}}$. As shown in Fig. 5(b), the optimal signal selected by the vocal energy estimator, after signal smoothing and temporal concatenation, contains more complete vocal components while reducing large-scale leakage caused by motion. Full implementation details, including the vocal-energy estimator, signal-smoothing procedure, and parameter settings, are provided in Appendix A.2.1.

3.1.2 *Body Motion Demodulation.*

By stitching optimal bins $\tilde{S}_{\text{opt}}$ across consecutive windows, we obtain a composite signal $\hat{\mathbf{S}}_{\text{opt}}$. But it retains residual motion interference due to the spatial extent of each bin. Within its phase differences, motion-induced components overlap with vocal vibration frequencies and introduce distortion (Fig. 5(a) and (b)).

Observations. Prior work [27, 87] generally assumes that body movement primarily affects the low-frequency region of mmWave, far below vocal fundamentals (e.g., 90 Hz [114]), and can therefore be removed by filtering. Yet our empirical observations contradict this assumption (Fig. 5(a) and (b)). Our theoretical derivation shows that the body motion frequency in phase differences scales with acceleration and cannot physiologically reach the vocal band. This motivates us to examine factors beyond motion, and our experiments reveal a previously underexplored mechanism: *the radar's zero-frequency DC component couples with body movements.* This coupling produces two observable effects:

(1) It generates pseudo-frequencies that cause overlapping distortion in the phase differences.

(2) Removing the DC component can demodulate the overlapping distortion in phase differences, but it introduces discontinuities when the DC overlaps with static reflections during stationary periods.

Complete derivations and experiments are provided in Appendix A.2.2. Building on this, we propose a motion demodulation algorithm that selectively removes the DC component during movement to eliminate overlapping distortion while preserving signal continuity during stationary periods, yielding refined vocal signals. The process is as follows.

Selective Removal of the DC Component. To identify the zero-frequency DC component and detect body movement, we operate in the time-frequency domain of $\hat{\mathbf{S}}_{\text{opt}}$, where frequency characterizes the Doppler velocity of reflective targets (see Appendix A.1). By analyzing the Doppler spectrogram obtained via Short-Time Fourier Transform (STFT) [85] (Fig. 6(a)), we identify a zero-frequency DC component and further reveal a dominant Doppler frequency corresponding to the torso's motion velocity, due to its large reflective surface. Using this Doppler cue, we design a motion-aware filtering method that, during STFT processing, applies a filter to remove the DC component only when the body's velocity exceeds a predefined threshold. Specifically, the $\hat{\mathbf{S}}_{\text{opt}}$ is segmented into frames, and the spectrum $P_i(f)$ is computed for each frame, where i is the frame index and f is the frequency. The motion-aware filtering is then implemented using a conditional smoothed band-stop filter, activated only when the spectral peak f_{peak} (torso motion) exceeds the threshold Δ, thereby suppressing near-zero frequencies while preserving static body reflections. The filter, referred to $H_{\text{bs}}(f)$, is designed with a stop band defined as $|f| \le \Delta$ and further incorporates a smoothed transition band to mitigate spectral leakage from sharp frequency cutoffs (See Appendix A.2.3). Each spectrum is finally given by:

$$\tilde{P}_i(f) = \begin{cases} P_i(f) \cdot H_{\text{bs}}(f), & |f_{\text{peak}}| > \Delta \\ P_i(f), & \text{otherwise} \end{cases}. \tag{1}$$

If $|f_{\text{peak}}| \le \Delta$, the frame is considered motionless, and low-frequency interference far below the vocal band is ignored.

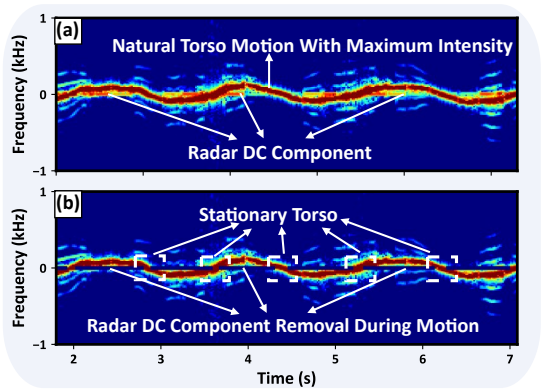

Figure 6: Motion demodulation in IQ signals. (a) Torso motion shows maximum intensity due to its large reflective area; when static, the DC component overlaps with torso reflections. (b) Motion demodulation removes the DC component during motion while preserving stationary torso reflections.

After the motion-aware filtering, the DC component is removed during motion but preserved during stationary periods, as shown in Fig. 6(b). We then apply Inverse Short-Time Fourier Transform (iSTFT) [85] to reconstruct the motion-demodulated mmWave IQ signal, with distortion in its phase difference effectively removed, as illustrated in Fig. 5(c). Finally, a high-pass filter is applied to suppress residual low-frequency clutter, as shown in Fig. 5(d).

3.2 Cross-transfer Pipeline

Compared with microphones, mmWave radar resists airborne noise, capturing vocal vibrations that align closely with voice emotion-related source features. However, collecting large-scale mmWave data is labor-intensive, while limited data leads to degraded performance and increases the risk of overfitting, partly due to incomplete harmonic capture of voice. To address this, we propose a cross-modal transfer pipeline that acts as a regularizer by distilling knowledge from an audio teacher to guide the mmWave student in emotion learning. Comprehensive audio representations improve offline convergence, while mmWave ensures noise-robust online inference. Specifically, we design a hybrid layer-wise distillation loss to mitigate performance degradation caused by the feature deficiency of mmWave.

3.2.1 Teacher Model Design. Representation learning is a powerful tool for training better supervised models using large-scale unlabeled data, especially when labeled data is scarce. Therefore, we adopt TRILL [102], a non-semantic speech embedding model pre-trained on AudioSet [34], which achieves state-of-the-art performance across a variety of non-semantic tasks. Building on TRILL, we fine-tune the model on collected emotional speech data by unfreezing its last two fully connected layers to learn task-specific representations. In addition, since word timing in speech correlates with emotional expression, we extract it via Google Cloud Speech-to-Text [38] as a complementary feature for emotion recognition. Notably, the calls for word timing are used only by the teacher model during offline distillation and are not involved in *FeelWave*'s online inference, thereby introducing no latency.

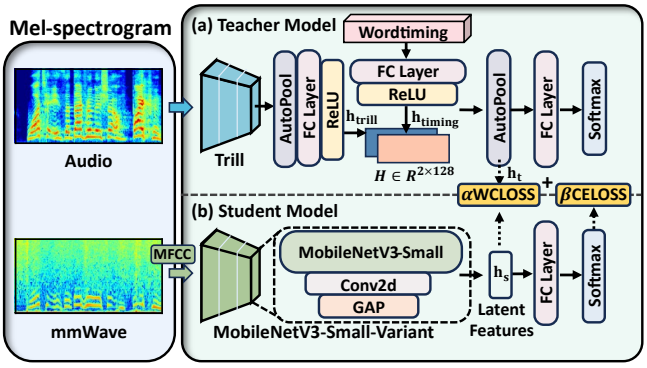

Figure 7: Cross-transfer pipeline.

Fig. 7(a) shows the architecture of the teacher model. TRILL, built on a modified ResNet-50 [45], extracts a variable-length embedding sequence $\{x_1, x_2, \ldots, x_T\}$, where $x_t \in \mathbb{R}^{2048}$. To handle variable-length speech inputs, we apply AutoPool [72] to aggregate the embedding sequence into a fixed-dimensional representation $\mathbf{z}_{\text{trill}} \in \mathbb{R}^{2048}$ by assigning adaptive attention weights to emphasize emotion-relevant frames. The aggregated representation is then passed into a fully connected layer with ReLU activation to obtain the emotion representation $\mathbf{h}_{\text{trill}} = \text{ReLU}(\mathbf{W}_1\mathbf{z}_{\text{trill}}+\mathbf{b}_1), \mathbf{h}_{\text{trill}} \in \mathbb{R}^{128}$. In parallel, a one-dimensional word timing prior $\mathbf{x}_{\text{timing}}$, encoding rhythm- and rate-related cues, is projected into the same space, yielding $\mathbf{h}_{\text{timing}} \in \mathbb{R}^{128}$. These two features are concatenated as $\mathbf{H} = \left[\mathbf{h}_{\text{trill}}; \mathbf{h}_{\text{timing}}\right], \mathbf{H} \in \mathbb{R}^{2\times128}$, and further adaptively fused using AutoPool to generate a unified emotion representation $\mathbf{h}_{\text{fused}} \in \mathbb{R}^{128}$. Finally, $\mathbf{h}_{\text{fused}}$ is fed into a softmax classifier to produce the emotion prediction:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_3\mathbf{h}_{\text{fused}} + \mathbf{b}_3), \quad \hat{\mathbf{y}} \in \mathbb{R}^C, \tag{2}$$

where C denotes the number of emotion categories. All fully connected layers employ L2 regularization (1×10^{-5}) to prevent overfitting and enhance generalization.

Comparative Experiments on Public Datasets. We evaluate the teacher model on four public emotion datasets: CASIA [122], EMODB [13], SAVEE [55], and RAVDESS [68], which share six emotions (angry, fear, happy, neutral, sad, surprise), with RAVDESS also including disgust and calm. As shown in Fig. 8, the teacher model surpasses two state-of-the-art baselines, CPAC [113] and Emotion2Vec [71], achieving average accuracy (ACC) gains of +13.84, +5.84, +4.38, and +3.26 percentage points on CASIA, EMODB, SAVEE, and RAVDESS under 10-fold cross-validation. Removing word timing input (W/O wordtiming) consistently reduces performance across all four datasets, with average drops of 2.34 percentage points in ACC and 2.81 percentage points in weighted F1 score (WF1). These results show that the teacher model captures emotional cues effectively and benefits from word timing information.

3.2.2 Training Scheme Design. Although the teacher model performs well under quiet conditions, its high noise sensitivity and large size (approximately 198 MB) limit its suitability for real-world deployment.

Noise Impact on Audio-based Emotion Recognition. The performance of the audio-based model degrades significantly under low SNRs. We record an emotion-labeled speech dataset with six

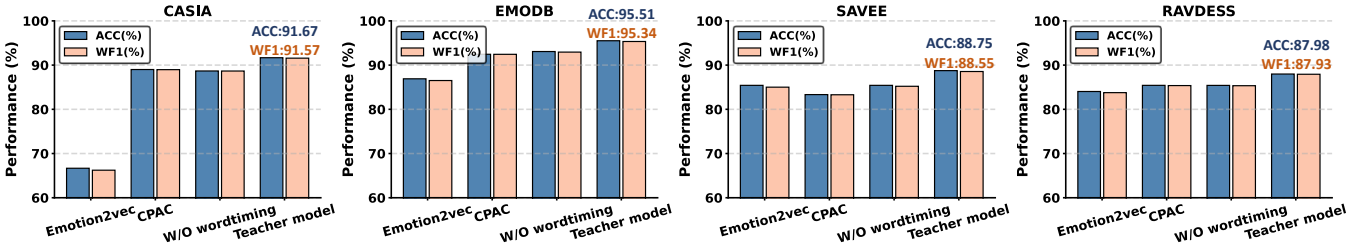

Figure 8: Comparison of accuracy (ACC) and weighted F1 score (WF1) across different methods and datasets.

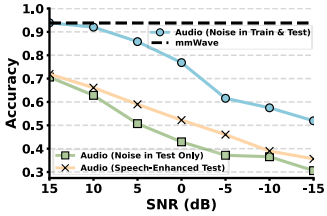

Figure 9: Impact of acoustic noise on audio models.

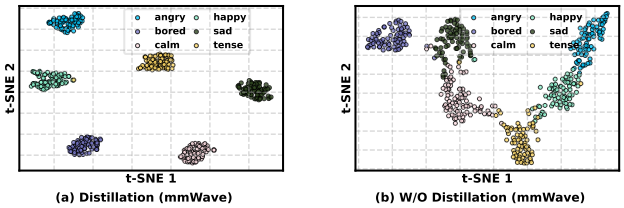

Figure 10: High-level feature distributions of the mmWave model with (a) and without (b) distillation.

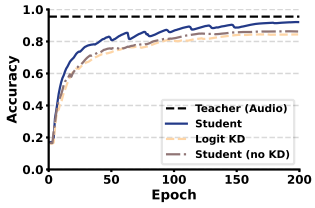

Figure 11: Validation accuracy on the mmWave dataset.

categories (happy, calm, angry, tense, sad, and bored) and simulate noisy environments by mixing the speech with natural background noise at SNRs ranging from 15 dB to -15 dB. Specifically, we evaluate the teacher model (W/O wordtiming) under these conditions. In addition, we apply Resemble-Enhance [95], a state-of-the-art speech enhancement toolkit, for noise suppression. As shown in Fig. 9, the accuracy of the audio-based teacher model drops substantially, even when the training set is adapted with the same type of noise. In real-world scenarios, however, diverse and unpredictable noise makes training adaptation infeasible. Additionally, after enhancement, its performance still remains poor, with the accuracy only 46.02% at -5 dB. The limitation persists since residual noise and spectral distortion after enhancement continue to hinder effective emotion recognition. In contrast, the mmWave modality remains insensitive to acoustic noise, offering a robust alternative to audio and a scalable direction for advancing emotion-aware agents.

Therefore, we design a cross-transfer pipeline to distill knowledge from audio to mmWave, providing lightweight and noise-robust emotion sensing while maintaining high performance. Fig. 7 illustrates the detailed training scheme. The student model takes 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) computed from the phase difference of mmWave IQ signals as input. Considering the low computational cost and suitability for mobile devices, we design the student model as a variant of the MobileNetV3-Small [47], followed by fully connected layers, with a total model size of 9.06 MB. Specifically, we reduce the final convolutional layer's channels to 128 and apply global average pooling to align its representation with the teacher model's fused feature vector. The teacher model is fine-tuned on our speech emotion dataset to learn task-specific representations, then frozen and used as a learning anchor during distillation.

3.2.3 Hybrid Layer-wise Loss. The incomplete harmonics in mmWave vocal signals make it prone to overfitting the stronger fundamental components. As shown in Fig. 10 and Fig. 11, this representation degradation reduces feature separability across emotions, leading to suboptimal convergence. Even with logit-based knowledge distillation (Logit KD) [46], performance remains limited by feature deficiency. To address this, we design a hybrid layer-wise loss that jointly leverages more generalizable latent features and class-level supervision, enabling the mmWave model to extract balanced features, reduce overfitting, and enhance emotion separability. It comprises two components described below.

Weighted Contrastive Loss. To bridge the representational gap between mmWave and audio modalities and enhance class boundary discrimination, we propose a weighted contrastive loss (WCLoss). It aligns latent features across modalities and assigns greater weights to the most confusing negative pairs, guiding the student model to better separate emotion features. As shown in Fig. 7, given latent features $\mathbf{h}_t, \mathbf{h}_s \in \mathbb{R}^{N\times 128}$ from the teacher and student models (N is the batch size), the WCLoss is defined as:

$$\mathcal{L}_{\text{WCL}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\text{sim}(\mathbf{h}_s^i, \mathbf{h}_t^i)}{\tau}\right)}{\sum_{j=1}^{N} w_{ij} \cdot \exp\left(\frac{\text{sim}(\mathbf{h}_s^i, \mathbf{h}_t^j)}{\tau}\right)}, \tag{3}$$

where $\mathbf{h}_s^i$ and $\mathbf{h}_t^i$ are the normalized features of the i-th pair, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature parameter. The weighting coefficient is defined as $w_{ij} = w$ if $j \in \mathcal{H}_i$, and $w_{ij} = 1$ otherwise, where $\mathcal{H}_i$ denotes the indices of the top-k hardest negative samples (from different emotion classes) for anchor i, selected based on similarity, and $w > 1$ assigns larger weights to these samples.

Cross Entropy Loss. To train the model for multi-class emotion recognition, we use manual truth labels as supervision and employ the cross entropy loss. Specifically, for an input mmWave sample $\mathbf{x}_m$ with its ground truth label $\mathbf{y}_m$, the predicted output is denoted as $\hat{\mathbf{y}}_m$. The cross entropy loss is defined as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_m(c) \log \hat{y}_m(c), \tag{4}$$

where C represents the set of emotion classes, and $y_m(c)$ denotes the label of class c.

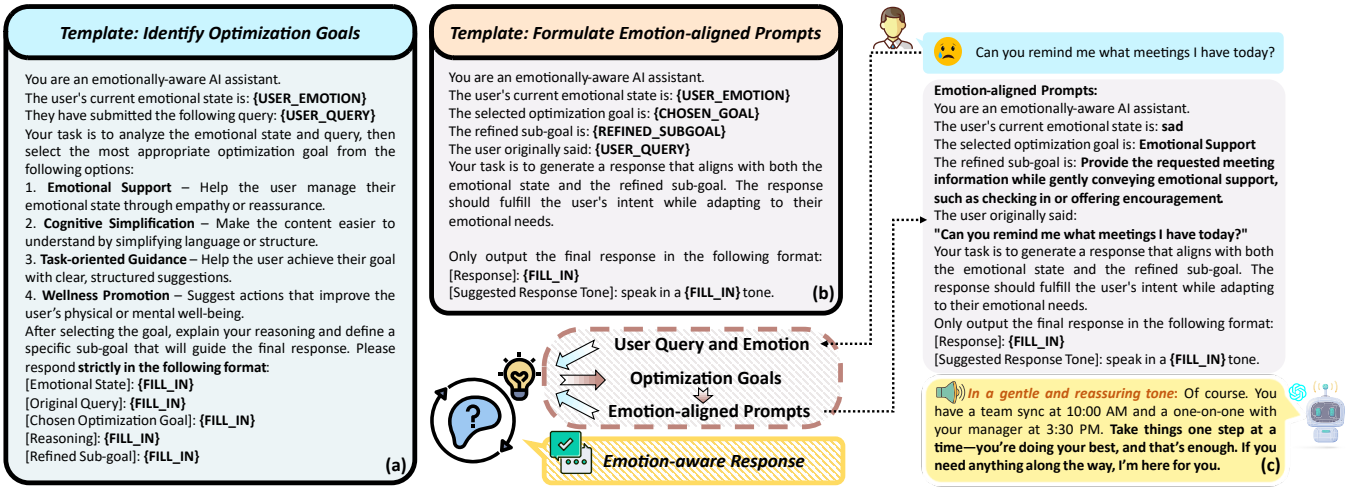

Figure 12: Emotion-driven interaction with GPT-4o. (a) Template for identifying the optimization goal; (b) Template for formulating the emotion-aligned prompt; (c) A running example.

Combining all components, the hybrid layer-wise loss is defined as $\mathcal{L}_{\text{hybrid}} = \alpha\mathcal{L}_{\text{WCL}} + \beta\mathcal{L}_{\text{CE}}$. As shown in Fig. 10, the cross-transfer pipeline yields clearer feature separation across emotions. Fig. 11 further shows that, compared with classic Logit KD without latent feature alignment, the proposed loss achieves improved performance, indicating emotion perception comparable to audio representations. This suggests that mmWave vocal signals inherently contain sufficient emotional cues, yet dominant discrepancies in fundamental and harmonic features lead to suboptimal convergence. Our method better balances the features in the mmWave student model, enabling more effective emotion inference.

3.3 Emotion-driven LLM Interaction

Since LLMs rely on retrieval and have limited contextual reasoning ability, insufficient reasoning from user queries and emotions to underlying needs often results in superficial responses [6, 91]. To this end, we present a two-stage emotion-driven query optimization module (QOM), which leverages users' emotional states to conduct an in-depth analysis of their needs and refine user queries. The goal is to enable the LLM agent to generate responses that are both contextually appropriate and aligned with the user's current emotional state. Specifically, we use the GPT-4o-transcribe [83] to obtain user speech queries, which has been proven effective under noisy conditions. The chain-of-thought (CoT) approach [63, 93, 112] has shown effectiveness in enabling LLMs to thoroughly analyze user context. Inspired by this, the QOM guides the LLM through intermediate reasoning steps that mimic human problem-solving, decomposing complex tasks into manageable sub-tasks for deeper, more contextually relevant responses. The process is as follows.

3.3.1 Identify the Optimization Goals. To overcome LLMs' limited reasoning over user queries and emotions, we model latent user needs as explicit, interpretable semantic representations, then employ retrieval-augmented generation (RAG) [28] to enhance contextual reasoning for accurate need inference. Grounded in the taxonomy of emotion-regulation motives [104], human emotions are closely intertwined with diverse psychological and functional needs, such as optimizing task performance, enhancing cognitive understanding, and maintaining mental and physical well-being. To enable the system to more effectively respond to users' underlying needs, we draw upon established principles from *emotional* [92], *cognitive* [103], *behavioral* [4], and *health psychology* [32] to design the following optimization goals.

- **Emotional Support**: Help the user manage their emotional state through empathy or reassurance.
- **Cognitive Simplification**: Make the content easier to understand by simplifying language or structure.
- **Task-oriented Guidance**: Help the user achieve their goal with clear, structured suggestions.
- **Wellness Promotion**: Suggest actions that improve the user's physical or mental well-being.

For example, when the system detects a *sad* or *angry* state with a frustration-related query (e.g., "I've been struggling with work lately"), it applies the *Emotional Support* by adopting an empathetic tone to ease negative emotions. When the user is *tense* and seeks clarification (e.g., "Could you explain this algorithm again?"), the system follows *Cognitive Simplification*, refining and structuring the response to help the user focus on key information. Meanwhile, *Task-oriented Guidance* and *Wellness Promotion* support goal execution and mental-physical balance. The former provides structured actions for task completion, while the latter offers gentle well-being prompts to alleviate stress from prolonged negative emotions. As shown in Fig. 12(a), we design a prompt template that reasons jointly over the user's emotional state and query context, guided by predefined optimization goals to derive refined sub-goals summarizing user needs.

3.3.2 Formulate the Emotion-aligned Prompt. Next, *FeelWave* performs structured prompting for the LLM. As shown in Fig. 12(b), we design a query template that integrates the user's query, emotional state, and sub-goals as guidance to fulfill the user's intent while adapting to their needs. *FeelWave* then reformulates the query to enhance the contextual relevance and empathy of the LLM's reply.

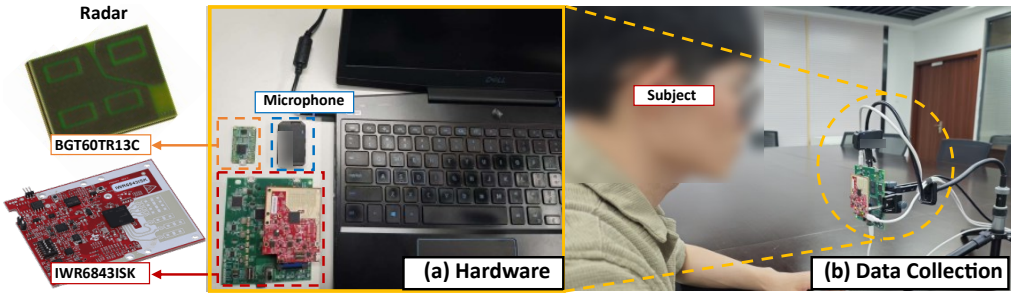

Figure 13: Data collection of *FeelWave*. (a) The hardware setup includes an mmWave radar and a microphone. The BGT60TR13C and IWR6843ISK are used for mobile and fixed devices, respectively; (b) A radar and a microphone are employed to collect paired data from subjects.

Fig. 12(c) presents a concrete example. *FeelWave* first analyzes the user's vocal vibrations and detects a sad state. Without the two-stage QOM, a GPT-4o agent responds to the user's query, "Can you remind me what meetings I have today?" with, "I'm sorry you're feeling sad... If you need anything else or some support, I'm here to help." Such a response provides surface-level empathy. In contrast, *FeelWave* performs deeper reasoning, selects the *Emotional Support* goal, and summarizes user needs as "Provide the requested meeting information while gently conveying emotional support, such as checking in or offering encouragement." Guided by this, emotion-aligned prompts are then passed to the agent, which generates a response that not only lists the user's meetings but also expresses deeper empathy as *"Take things one step at a time—you're doing your best, and that's enough,"* delivering both functional and emotional support in a gentle, reassuring tone. Notably, while *FeelWave* explicitly leverages mmWave vocal vibrations for noise-robust emotion analysis, it does not discard semantics. Instead, it integrates the LLM's implicit reasoning in the QOM to infer contextual and latent emotional cues from user queries. By combining dominant explicit affect derived from mmWave sensing with auxiliary implicit affect inferred through semantic reasoning, *FeelWave* jointly optimizes the LLM's responses and avoids emotional cue bias across modalities in a cost-effective strategy, without LLM fine-tuning.

4 EVALUATION

We focus on how emotional integration enhances user experience during voice interactions with intelligent assistants. To this end, we conduct preliminary experiments to **(1)** examine the impact of user emotional states on agent response quality and **(2)** evaluate *FeelWave*'s emotion sensing performance. All our studies were approved by the Institutional Review Board (IRB) of our institution.

4.1 Experimental Setup

Hardware. As shown in Fig. 13, we use two types of commercial mmWave radars, TI IWR6843ISK [52] and Infineon BGT60TR13C [50], to capture vocal vibration signals. Both devices transmit FMCW signals in the 60-64 GHz range. *FeelWave* employs a radar in a single-input single-output (SISO) configuration with one transmit and one receive antenna. The radar operates at 3.6 GHz bandwidth and 10 kHz sampling rate to ensure complete capture of vocal vibrations. Additionally, the SISO radar, with a horizontal and vertical field of view (FOV) of approximately 40°, can robustly perceive users' emotions through vocal vibrations, meeting the spatial requirements of typical real-world device placements. The detailed configuration is provided in Table 1. To support portability in mobile scenarios, we adopt the BGT60TR13C, while the IWR6843ISK is used in fixed settings.

Software. Leveraging open-source APIs, including the real-time mmWave recording tool [120] and ifxdaq [49], we develop Python scripts to control the TI IWR6843ISK and Infineon BGT60TR13C radars for capturing vocal vibration signals. *FeelWave* is trained via cross-domain transfer on a server with an NVIDIA GeForce RTX 4090 GPU using TensorFlow 2.15.0. For deployment, we port the mmWave model (2.38 M parameters, 9.06 MB) to a laptop with an AMD Ryzen 9 HX 370 CPU, enabling real-time processing. For a 5-second data segment, *FeelWave* requires 21.71 ms (± 0.11 ms) for preprocessing. The inference latency is 5.46 ms (± 0.32 ms) on the GPU and 25.69 ms (± 1.86 ms) on the CPU. This yields a real-time factor (RTF) below 1, satisfying real-time requirements. To enable online voice interaction, we integrate GPT-4o [78] into *FeelWave* and develop an app, as illustrated in Fig. 21. We use GPT-4o-transcribe [83] to extract clean user queries, which has demonstrated state-of-the-art performance under noisy conditions [65, 80]. Finally, we use GPT-4o-mini-TTS [82] to synthesize voice responses. We also evaluate GPT-4o's interaction latency through time-to-first-token (TTFT) and time per output token (TPOT). TTFT measures the delay from query reception to the first token, while TPOT measures the average generation time per token. Across five trials, TTFT averages 0.822 s while TPOT averages 0.001 s/token. Although query optimization introduces a slight delay before response generation, GPT-4o maintains efficient token generation once decoding begins.

Dataset. Over one month, we recruited 27 stage actors (11 male, 16 female, aged 18-31) to perform emotional enactments while collecting synchronized mmWave and audio data, forming the *EmoDataset*. We adopted a six-category emotion taxonomy comprising *happy*, *calm*, *angry*, *tense*, *sad*, and *bored*, covering a broad range of valence and arousal for emotional speech. Each participant selected 6-9 spoken commands from ok-google.io [37] and performed them

Table 1: mmWave radar configuration.

Frame Periodicity	1 ms	**Idle Time**	40 μs
Chirps/Frame	10	**Ramp End Time**	60 μs
Frequency Slope	60 MHz/μs	**Range Resolution**	4.2 cm

under all six emotional states. During recording, participants spoke naturally while maintaining a distance of 0.3-0.5 m from the radar. The *EmoDataset* contains roughly 12 hours of recordings.

4.1.1 User Query Design. To systematically evaluate the impact of emotional state on LLM-powered agent responses, we design three query types corresponding to common user intents: information retrieval ("What is it"), procedural guidance ("How to do"), and task execution ("Please do..."). These categories reflect three core user intent types: declarative, procedural, and imperative. For each type, we construct multiple representative scenarios (see Appendix A.3) covering common tasks such as checking flights, requesting navigation, and playing music. This design offers clear semantic cues for the agent, enabling evaluation of how emotional context impacts response quality, as detailed below.

- **Information retrieval queries** assess whether emotional cues enable the agent to convey factual content more accurately or empathetically (e.g., offering gentler replies during sadness).
- **Procedural guidance queries** examine whether the agent adapts its instructional language based on emotional states (e.g., offering clearer, more concise instructions under anxiety).
- **Task execution queries** evaluate whether the agent incorporates emotional states when executing commands to provide more humanized responses or confirmations (e.g., playing upbeat music when happy).

This design enables a systematic analysis of how emotional conditioning affects language generation across interaction types, clarifying how emotion shapes agent behavior.

4.1.2 User Study Design. We design a questionnaire to examine how emotional cues affect user interaction with the agent. The questionnaire includes (1) the user's emotional state inferred from mmWave-based vocal vibration sensing, (2) the user's query and its context, and (3) the response generated by the LLM-powered agent. For each query, participants receive both a baseline agent response without emotional state and an emotion-enhanced response, and are asked to indicate their preference. To mitigate order effects, the presentation order of the responses is randomized. Participants rate each response on a 5-point Likert scale across the following five dimensions:

- **Clarity**: how easy it is for the user to understand the response.
- **Emotional Appropriateness**: whether the response aligns with the user's emotional state and appropriately addresses emotional needs.
- **Contextual Relevance**: whether the response is relevant to the user's request and the current context.
- **Interaction Comfort**: whether the response makes the user feel more relaxed and more willing to continue.
- **Overall Satisfaction**: the user's overall satisfaction with the response.

Additionally, we evaluate users' preferences between the two responses for each dimension to determine which one is favored. The options are: (1) Prefer response 1, (2) Prefer response 2, (3) Like both, and (4) Dislike both.

4.2 Preliminary Experiments

In this section, we design preliminary experiments to (1) simulate users to broadly evaluate changes in user experience after introducing emotional states, and (2) verify the accuracy of *FeelWave* in recognizing users' emotions.

4.2.1 User Simulation Study. We conduct a user simulation study involving 20 human participants and five LLMs: GPT-4o [78], Gemini 2.5 Pro [22], Claude 3.5 Haiku [7], Llama 3.1 [75], and DeepSeek V3 [23], referred to as *LLM participants*. Human participants were recruited via public calls on social media, comprising 11 females and 9 males aged 18-35. Following recent methodological practices that use LLMs as simulated participants to emulate user feedback in controlled, reproducible settings [61, 89], we employ these models as complementary evaluators that provide consistent, bias-reduced assessments, thereby balancing human variability in subjective evaluation. Using multiple LLMs further reduces single-model bias and broadens cross-model perspectives. Combining both participant types enables a more comprehensive view of how emotional cues shape perceived interaction quality. Each *LLM participant* provides 4 repeated ratings to match the number of questionnaire responses from human participants. We design three user inquiry categories: information retrieval (T1), procedural guidance (T2), and task execution (T3), as described in Section 4.1.1. Based on this, we simulate diverse dialogues across 36 queries, comprising 6 commands expressed under 6 emotional states. For example, in the inquiry "Do I need an umbrella for tomorrow?", the user plans outdoor activities while feeling tense. We use a GPT-4o [78] agent to generate responses. As detailed in Section 4.1.2, participants view both a neutral response and an emotion-enhanced response generated with the emotion-aligned prompt (Fig. 12), then evaluate them using predefined metrics and indicate their preference.

Results analysis. We systematically evaluate the impact of emotional cues on user experience using both human participants and *LLM participants* as simulated users. As shown in Fig. 14 and Fig. 15, emotional cues consistently improve most metrics across both groups, with notably strong gains in emotional appropriateness and interaction comfort. In terms of overall preference, Fig. 16 shows that 84.7% of all participants favor emotionally enriched responses, underscoring their clear benefit in enhancing user experience. To further validate these findings, we conduct Wilcoxon signed-rank tests and estimate effect sizes with Cohen's d. As shown in Table 2, emotional cues significantly enhance overall satisfaction for participants as a whole, including both humans and LLMs ($p < 0.05$), with most effect sizes in the medium to large range. By task type, emotionally enriched responses perform best in T2 and T3, significantly enhancing interaction comfort and overall satisfaction. In T1, user experience also improves, though clarity scores are slightly lower with emotional cues (4.46 ± 0.89 vs. 4.56 ± 0.89). These results suggest that in information retrieval tasks, users favor direct and concise responses, while emotional embellishments, particularly under angry or tense states, may slightly reduce perceived clarity. In contrast, emotional cues help the agent provide appropriately detailed responses in procedural guidance (T2), yielding a significant advantage in clarity. For both *LLM participants* and human participants, the results show a consistent trend. The former provide a more stringent and objective perspective, while the latter offer

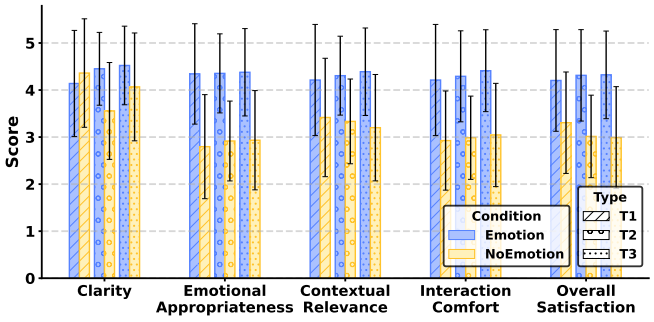

Figure 14: Human participant simulation results.

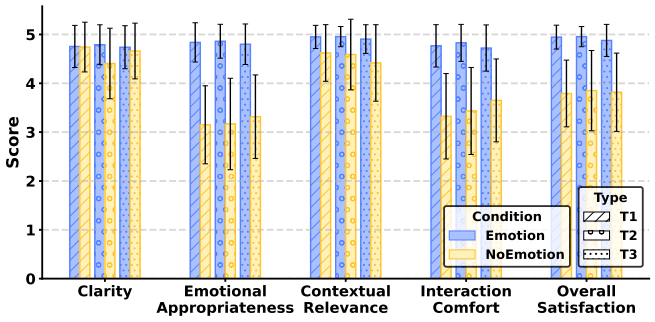

Figure 15: LLM participant simulation results.

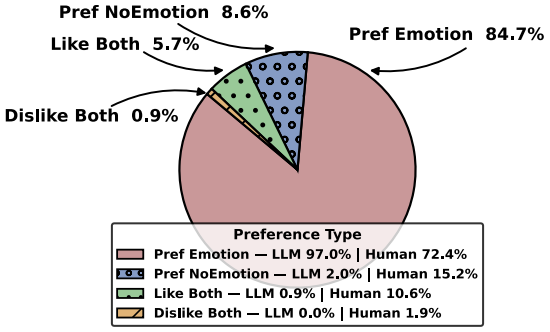

Figure 16: Participant preferences for agent responses with and without emotional states.

authentic subjective experience, and both indicate that emotional cues reliably enhance agent response quality. The corresponding statistical analyses for each group are reported in Appendix A.4. Overall, incorporating users' emotional states enhances the interaction experience across multiple dimensions.

Note. In the early recruitment phase, we also contacted potential participants from older age groups. Many reported limited prior exposure to LLMs or agent-based systems, making them less familiar with the interaction tasks and unable to provide stable or comparable feedback on interaction quality. To avoid confounding effects due to differences in technology familiarity and to ensure that evaluations reflect actual usage experience, we ultimately focused on younger users who more frequently engage with intelligent agents and represent the primary target user group for such systems.

4.2.2 mmWave-Based Emotion Recognition Performance. To evaluate the emotion sensing performance of *FeelWave*, we conduct experimental analysis and compare it against state-of-the-art approaches. To ensure fairness, we conduct 5-fold cross-validation on the *EmoDataset*, which includes 27 participants evenly spanning 6 emotional states.

Baselines. We adopt the following cross-domain learning methods as baselines: **(1)** Logit-based Knowledge Distillation (Logit KD) [46], a classical distillation approach without latent feature alignment; **(2)** Adversarial Teacher-Student Distillation (ATSD), which applies a gradient reversal layer (GRL) [33] to adversarially align teacher-student latent representations across modalities. All frameworks share the same network architecture and differ only in distillation strategy. We further include **(3)** CPAC [113], a state-of-the-art speech emotion recognition model trained directly on mmWave data without distillation, as well as two boundary baselines: **(4)** a non-distilled mmWave student model as the lower bound, and **(5)** an audio-based teacher model as the upper bound.

Overall performance. We evaluate *FeelWave* using average accuracy (ACC) and weighted F1 score (WF1) against multiple baselines. As shown in Fig. 17, *FeelWave* achieves the best overall results, with average gains of about 5.3 percentage points in both ACC and WF1. Notably, the limited performance gap between non-distillation methods and distillation baselines suggests that while mmWave vocal signals inherently contain rich emotion-relevant cues, their utilization is constrained by an imbalanced representation, where dominant fundamentals overshadow harmonic patterns. Our method more effectively unlocks this latent potential, enabling more discriminative emotion representations. We also use the silhouette score (-1 to 1, higher is better) to assess the clustering quality of emotion features based on intra- and inter-class distances. For reference, we also compute the silhouette score of the teacher model with audio input. As shown in Fig. 18, *FeelWave* achieves

Table 2: Cohen's d and statistical significance of combined human-LLM evaluations across metrics and query types.

Metric	T1	T2	T3	All
Clarity	-0.16*	**0.70** ✓	**0.39** ✓	0.28
Emotional Appropriateness	**1.90** ✓	**1.79** ✓	**1.65** ✓	**1.78** ✓
Contextual Relevance	**0.71** ✓	**0.82** ✓	**0.96** ✓	**0.81** ✓
Interaction Comfort	**1.39** ✓	**1.39** ✓	**1.31** ✓	**1.36** ✓
Overall Satisfaction	**1.37** ✓	**1.45** ✓	**1.49** ✓	**1.42** ✓

Note: All results are statistically significant at $p < 0.05$. Each cell reports Cohen's d effect size, in bold with ✓ marking $d \geq 0.36$ (medium or large). According to Lovakov and Agadullina [69]: $d < 0.15$ = very small, 0.15-0.36 = small, 0.36-0.65 = medium, $d > 0.65$ = large. A medium effect size can be interpreted as noticeable to the observer.

* **The average score with emotional cues (4.46) is slightly lower than without emotional cues (4.56).**

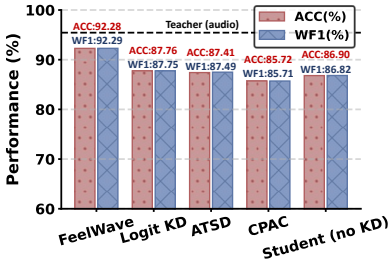

Figure 17: Performance of emotion recognition across five-fold cross-validation.

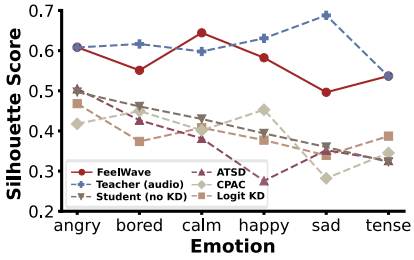

Figure 18: Silhouette score evaluation of emotion clustering performance.

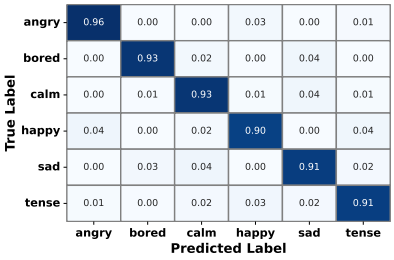

Figure 19: Confusion matrix of *FeelWave* on emotion recognition.

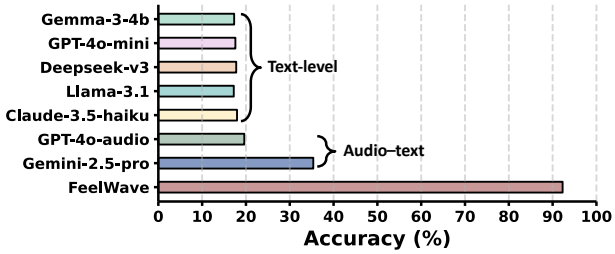

Figure 20: Performance of semantic-oriented models on emotion recognition.

the highest average score of 0.57 with mmWave input, surpassing Logit KD, ATSD, and the non-distillation baselines. Moreover, the silhouette score of *FeelWave* closely approaches that of the teacher model with audio input (0.61), indicating near-lossless feature discriminability. This demonstrates that our cross-transfer approach with the layer-wise hybrid loss more effectively guides the audio-to-mmWave latent feature mapping and reduces overfitting, leading to improved performance. As shown in Fig. 19, *FeelWave* robustly recognizes multiple emotions, maintaining ACC values above 90% across all categories.

Semantic-oriented emotional cue analysis. We introduce several semantic-oriented baselines to assess the role of semantic emotional cues in functional agent interactions. We assess five text-level models, including Claude 3.5 Haiku [7], Llama 3.1 [75], DeepSeek V3 [23], GPT-4o-mini [81], and Gemma-3-4B [21]. We also include two audio-text LLMs, Gemini 2.5 Pro [22] and GPT-4o audio [79], which leverage both semantic and acoustic cues. As shown in Fig. 20, both categories exhibit limited performance, with average ACCs of 17.57% and 27.60%, respectively. A primary reason is that functional queries (e.g., weather checks, schedule reminders) contain largely neutral semantics, making it difficult for text-level models to derive enough emotional cues. Audio-text models also suffer from bias between semantic and acoustic emotional cues. In functional interaction scenarios, semantic content is often neutral or conveys only coarse polarity (such as positive or negative), whereas acoustic signals carry more fine-grained emotional information. Prior studies [15, 19] indicate that current audio-text LLMs tend to prioritize semantics under such inconsistency, leading them to underuse acoustic affect and limiting effective cross-modal integration. Consequently, these models struggle to unify the emotion label space, which limits their overall performance. For example, GPT-4o-audio reaches only 19.62%. In contrast, *FeelWave* explicitly leverages mmWave-captured vocal features for noise-robust emotion analysis and incorporates LLM reasoning during query optimization to extract contextual and implicit emotional cues. This design combines the dominant explicit affect derived from mmWave sensing with auxiliary implicit affect inferred from semantic reasoning, avoiding cross-modal emotional cue bias and improving the quality of agent responses.

Robustness analysis. We comprehensively evaluate the generalization of *FeelWave* under diverse conditions, including unseen speakers, variations in radar distance and orientation, body movements, and everyday clothing occlusion. *FeelWave* generalizes well to unseen speakers, achieving an average ACC of 82.78% and up to 91.67% for emotion recognition. In addition, within a distance of 1.0 m, *FeelWave* supports both mobile scenarios (0-0.5 m, e.g., smartphones) and fixed-device scenarios (0.5-1.0 m, e.g., laptops, in-vehicle systems). It is also robust to user-radar orientation, including Frontal, Top-left, and Bottom-left, covering azimuths up to approximately 60°, elevations from approximately +15° to -45°, and a horizontal and vertical FOV of around 40°. Furthermore, *FeelWave* remains robust against everyday movements (e.g., writing, typing, walking) and clothing occlusion of the vocal cord region (e.g., scarf, down jacket, jacket). Overall, *FeelWave* can be effectively applied to natural voice interaction across diverse device settings, such as smartphones, laptops, and in-vehicle systems. Detailed robustness results are provided in Appendix A.5.

5 USER STUDY ON VOICE INTERACTION WITH FEELWAVE

In this section, we present real-world user studies of voice interaction with *FeelWave*, conducted under institutional IRB approval, to evaluate its effectiveness in diverse real-time scenarios and to ground the findings in authentic human experiences. As shown in Fig. 21, we implement an app that integrates GPT-4o into *FeelWave*, forming an emotion-aware voice agent for online queries. The pipeline incorporates GPT-4o-transcribe for automatic speech recognition and GPT-4o-mini for text-to-speech, enabling accurate transcription and natural voice responses. For comparison, we also evaluate a baseline voice agent without emotion sensing.

5.1 Controlled User Study

We recruit 20 participants via public social media calls (12 females, 8 males, aged 18-30) to evaluate *FeelWave*. Participants engage in real-time voice interactions with both *FeelWave* and a baseline voice agent without emotion awareness. The experiment allows users to make open-ended queries. Fig. 22(a) and Fig. 23 show that users consistently prefer emotion-aware interactions. Compared with

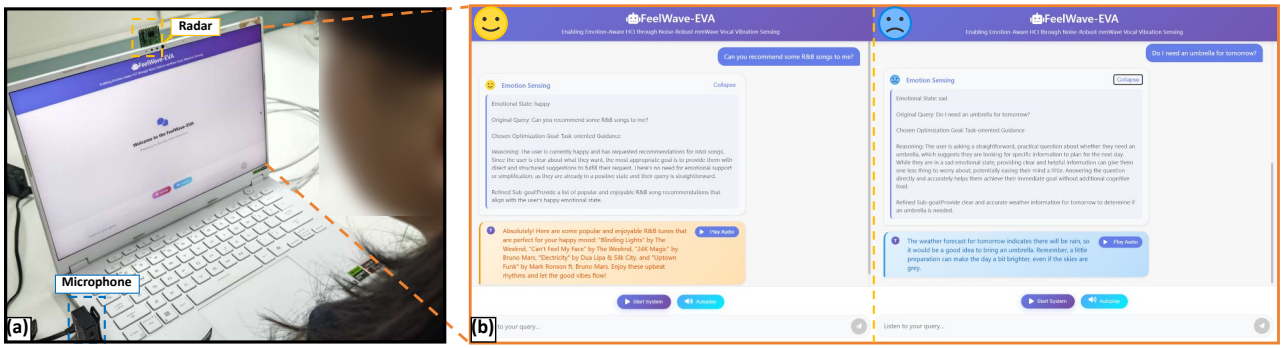

Figure 21: Voice interaction with *FeelWave*. (a) User study setup. (b) Example of the voice interaction app in use.

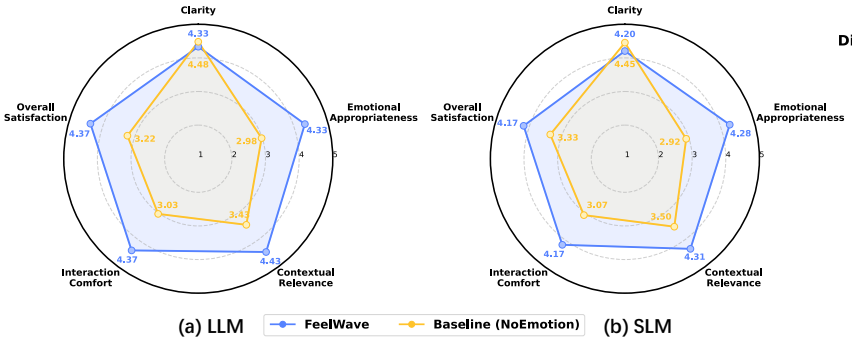

Figure 22: Results of voice interactions using *FeelWave*. (a) *FeelWave* using GPT-4o; (b) *FeelWave* using SLMs, including gpt-4o-mini and gemma-3-4B.

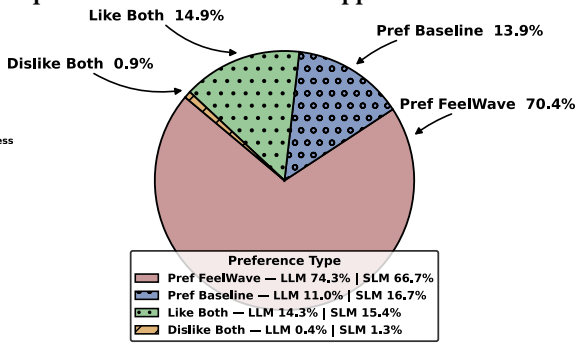

Figure 23: User preferences for voice interactions.

the baseline, *FeelWave* receives significantly higher ratings across emotional appropriateness (4.33 ± 1.31 vs. 2.98 ± 0.99), contextual relevance (4.43 ± 1.25 vs. 3.43 ± 1.25), interaction comfort (4.37 ± 1.27 vs. 3.03 ± 1.01), and overall satisfaction (4.37 ± 1.23 vs. 3.22 ± 1.03). Preference distribution analysis further reveals that with GPT-4o, 74.3% of interactions are judged more favorable when enhanced with emotional awareness.

Would model size matter? Currently, small language models (SLMs) can be deployed on mobile devices due to their lightweight nature, enabling on-device processing that preserves privacy while reducing computational costs. However, since model intelligence typically scales with parameter size [1, 111], a natural question arises: *Does using smaller models diminish the benefits of emotion-aware interaction?* To explore this, we replace GPT-4o in *FeelWave* with two SLMs, gpt-4o-mini [81] and gemma-3-4B [21]. As shown in Fig. 22 and Fig. 23, emotion awareness significantly enhances user experience regardless of whether an LLM or SLM is used, and the overall trend remains consistent. For example, under the SLMs, emotional appropriateness increases from 2.92 ± 0.97 to 4.28 ± 1.34, contextual relevance from 3.50 ± 1.21 to 4.31 ± 1.31, interaction comfort from 3.07 ± 1.00 to 4.17 ± 1.42, and overall satisfaction from 3.33 ± 0.97 to 4.17 ± 1.42.

Table 3: Cohen's d and statistical significance of *FeelWave*'s results.

Metric	LLM	SLM	All
Clarity	-0.19†	-0.28*	-0.24*
Emotional Appropriateness	**1.33 ✓**	**0.96 ✓**	**1.09 ✓**
Contextual Relevance	**0.94 ✓**	**0.67 ✓**	**0.79 ✓**
Interaction Comfort	**1.36 ✓**	**0.80 ✓**	**1.01 ✓**
Overall Satisfaction	**1.19 ✓**	**0.61 ✓**	**0.83 ✓**

Note: Each cell reports Cohen's d effect size. Superscript † indicates $p \geq 0.05$ (not significant). Values in bold with ✓ mark significant effects with $d \geq 0.36$ (medium or large). A superscript * denotes significant cases where the baseline outperformed *FeelWave*.

Overall, as shown in Table 3, both the Wilcoxon signed-rank and Cohen's d confirm that *FeelWave* consistently enhances emotion-aware interactions under both LLM and SLM settings. With GPT-4o, large effect sizes are observed in emotional appropriateness ($d = 1.33$), interaction comfort ($d = 1.36$), and overall satisfaction ($d = 1.19$). Although smaller models yield slightly lower scores, the improvements remain significant with medium-to-large effect sizes. Although some participants note reduced clarity in information-retrieval queries due to emotional modulation, this effect does not outweigh the overall benefits. Specifically, participants remarked:

> **Query**: *"Do I need an umbrella for tomorrow?"* ***(tense)***
>
> ***FeelWave***: *"It looks like there's some rain in the forecast for tomorrow, so bringing an umbrella would be a smart choice. Don't worry—we'll make sure you're prepared so your plans can still go smoothly."*
>
> **Baseline**: *"Yes, you'll need an umbrella—rain is in the forecast for tomorrow."*
>
> **Comment 1** :*"I feel the FeelWave response is warmer and more caring. It makes me feel less anxious, while the baseline answer sounds a bit cold."*
>
> **Comment 2** :*"When tense, I prefer brief replies. I like that FeelWave gives me the answer first and then adds comfort, which makes the longer response acceptable."*

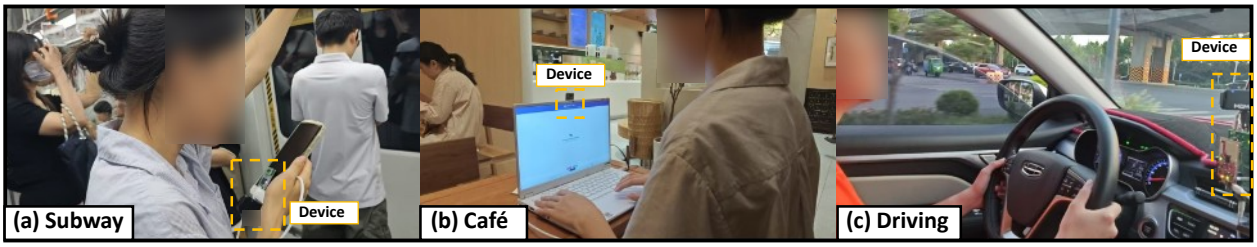

Figure 24: Real-world use cases. (a) Subway scenario with noise levels around 75 dB (67-79 dB); (b) Café scenario with noise levels around 58 dB (39-70 dB); (c) Driving scenario with noise levels around 73 dB (64-77 dB).

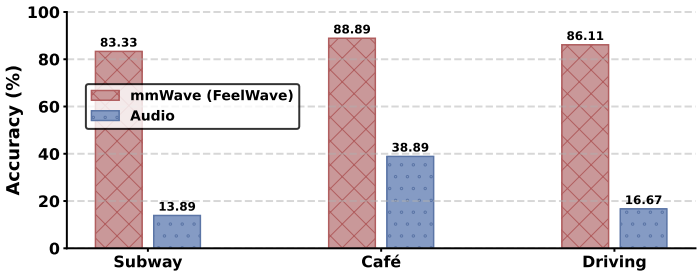

Figure 25: Emotion recognition performance of *FeelWave* in real-world conditions.

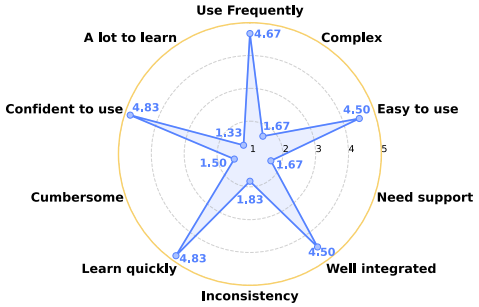

Figure 26: System Usability Scale (SUS) results from user evaluation.

Thus, LLMs deliver stronger performance, but SLMs still reliably reproduce the gains, supporting *FeelWave*'s feasibility for resource-constrained edge deployment.

5.2 In-the-Wild Study

To further validate *FeelWave* in real-world conditions with strong noise and natural movement, we conduct a voice interaction study *in a subway, a café, and a driving scenario*. As shown in Fig. 24, users interact in the subway, standing and holding a device in a moving carriage, where ambient noise averages 75 dB (67-79 dB). In the café, the radar is mounted above the laptop camera, with users interacting naturally in ambient noise averaging 58 dB (39-70 dB). In driving, the radar is mounted above the dashboard as users drive, introducing fine-grained vibrations and natural movement under noise averaging 73 dB (64-77 dB). We evaluate the emotion recognition performance of *FeelWave* in the above scenarios. In addition, we employ the System Usability Scale (SUS) [12] to assess users' voice interactions with *FeelWave*, focusing on dimensions such as ease of use, complexity, and user confidence. SUS consists of 10 standardized items rated on a five-point Likert scale and is widely used to assess system usability. The overall score is computed as:

$$\text{SUS} = 2.5 \times \sum_{i=1}^{5} \Big[(R_{2i-1} - 1) + (5 - R_{2i}) \Big], \tag{5}$$

where R_j denotes the raw rating of the j-th item, yielding a final score between 0 and 100.

As shown in Fig. 25, *FeelWave* achieves an average emotion recognition accuracy of 86.11% in noisy real-world settings such as subways, cafés, and driving scenarios with natural motion. In contrast, the audio-based teacher model achieves only 23.2% average accuracy under noise, while *FeelWave* remains effective across diverse conditions by leveraging mmWave's resilience. After using *FeelWave* for voice interactions, users evaluate its usability. The SUS results (Fig. 26) indicate that users consider the system suitable for frequent use (4.67 ± 0.47), easy to use (4.50 ± 0.50), and well-integrated (4.50 ± 0.50), with emotion sensing and goal optimization enhancing interaction effectiveness. Its transparent design enables natural voice interaction without additional effort, yielding a high learnability score (4.83 ± 0.37), and users express strong confidence in using the system (4.83 ± 0.37). Overall, *FeelWave* achieves a SUS score of 88.3, indicating excellent usability (≥ 85 is considered "excellent" [11]).

Feedback. In real-world use of *FeelWave*, users mainly raised concerns about safety, privacy, robustness under atypical vocal conditions, and cross-scenario applicability. The feedback and responses are summarized in Table 4. These insights highlight the system's usability strengths while also pointing to areas for refinement. Specifically, we further discuss the system's limitations and directions for future improvement in Section 6.3.

Exploratory Interview. This interview aims to examine the potential of *FeelWave* across diverse user backgrounds (e.g., daily routines, emotional needs, and familiarity with AI interactions). To this end, we recruit participants through public calls on social media, covering 4 students, an office worker, a taxi driver, a professor, and a retired senior, spanning ages 23 to 67. Interview details are provided in Appendix A.6. The interviews indicate that *FeelWave* offers promising opportunities for everyday voice interaction, such as stress regulation (e.g., streamlined scheduling and driving support), emotion-aware entertainment (e.g., personalized music playback), and lightweight companionship. At the same time, as noted by the professor, extending *FeelWave* from a personal agent to multi-user scenarios is appealing but would require further development of reliable user identification and multi-user dialogue management. Likewise, the retired senior views *FeelWave* as not yet a "friend-like" conversational partner, revealing a gap between emotional interaction and meaningful companionship. This suggests a

need for richer personalization, such as incorporating individual life experiences and long-term preferences, rather than relying on a generic assistant behavior. Overall, these insights position *FeelWave* as a promising design space, while also pointing to the need for improved extensibility to support more diverse applications.

6 DISCUSSION

6.1 Use Cases

Voice has become a mainstream interaction modality, and LLMs have further enabled more natural and intelligent conversations [8, 9]. Experiments show that *FeelWave* robustly senses user emotions within 1 m, across orientations, under clothing occlusion, and during natural movements, while remaining resilient to acoustic noise. User studies further confirm its ability to support personalized voice interactions in real-world scenarios. Specifically, the system is well-suited for two major categories of *open-environment voice interaction*, as follows.

6.1.1 Voice Interaction with Mobile Devices. *FeelWave* is well-suited for voice interaction with mobile devices such as smartphones and smartwatches. When users hurriedly query their assistant while rushing to catch a subway, traditional microphone-based systems are easily disrupted by strong background noise, often failing to capture emotional states and sometimes even producing errors that worsen negative emotions. In contrast, *FeelWave* leverages mmWave-based vibration sensing to unobtrusively perceive user emotions, enabling noise-robust, emotionally aligned interactions that can comfort users who may already feel physically and mentally fatigued. Overall, *FeelWave* is effective within typical mobile interaction ranges (within 0.5 m) and ±30° angles, supporting emotion-aware and reassuring voice interaction even in noisy public environments such as streets and shopping malls.

6.1.2 Voice Interaction with Fixed Devices. *FeelWave* can also be applied to fixed-device voice interactions, such as in-vehicle assistants and laptops. It can be deployed on a car dashboard or above a laptop's camera, typically within 1 m of the user. In driving scenarios, road rage often contributes to unsafe behavior. An in-vehicle assistant equipped with *FeelWave* can detect negative emotions such as anger from users' voice queries and provide calming feedback or timely reminders, thereby fostering a safer driving environment. Similarly, many office workers require mobile office support. When integrated into laptops, *FeelWave* enables robust voice interaction even in noisy environments such as offices or cafés, while adapting to users' emotional states to deliver more personalized and efficient workplace experiences.

6.2 Deployment Feasibility

6.2.1 Hardware Cost. Advances in mmWave technology have substantially reduced radar costs, enabling large-scale deployment. Consumer-grade radars such as the ICLEGEND MICRO cost approximately $2 [76], while Infineon and TI modules are priced around $12 [51] and $15 [53]. This demonstrates that integrating mmWave radar into *FeelWave* is cost-effective and lowers barriers to adoption.

6.2.2 Power Consumption. *FeelWave* uses a SISO mmWave radar with a low-power MEMS microphone. Smartphone-grade MEMS microphones consume about 1.8 mW [86], while the TI IWR6843ISK radar transmits at 12 dBm (≈ 16 mW) and consumes on average 1.2 mW with a 7.3% duty cycle. The Infineon BGT60TR13C radar consumes under 5 mW. These specifications show that *FeelWave* fits well within the energy budgets of power-constrained devices.

6.2.3 Integration Trends. Compact mmWave radars are increasingly embedded in consumer devices for interaction [41, 42, 94] and communication [99]. Smartphones such as Google Pixel 4 [94] already integrate radar sensors for gesture-based interaction, which demonstrates feasibility under mobile constraints. Similarly, wearable and portable devices (e.g., PieX Pendant [42], AIBI Pocket Pet [41]) combine low-power radars with edge AI processors, while in-vehicle systems use mmWave sensors for driver monitoring [10] and user interaction [77]. These developments indicate a convergence of radar sensing and edge intelligence, enabling cost-effective, low-power, privacy-preserving deployment. Within this trend, *FeelWave* is well aligned with future consumer applications.

6.3 Limitations and Future Work

In this paper, we propose an emotion-aware voice interaction system that operates effectively within a 1 m range, across multiple orientations, with clothing occlusion, and during natural body movements. It is suitable for mobile devices (e.g., smartphones) and fixed platforms (e.g., laptops, in-vehicle systems). Nonetheless, limitations remain in complex real-world environments, which we discuss here along with future directions for improvement.

Table 4: User feedback and responses.

Feedback	Response
"Does prolonged exposure to mmWave radiation pose health risks?"	mmWave is non-ionizing radiation and does not damage DNA. International safety guidelines (e.g., ICNIRP [54], FCC [31]) set strict exposure limits, and our devices operate at milliwatt-level power, far below these thresholds. Therefore, no health risks are associated with their use.
"Will this system expose my private information, such as my emotional states?"	Our system currently uses a hybrid local-cloud architecture: emotion analysis runs on-device to preserve privacy, while voice interaction via the OpenAI API is processed in the cloud in compliance with data protection standards [84]. Alternatively, it can be fully localized using open-source models (e.g., the Gemma family [21]).
"Will recognition accuracy decrease when my voice is hoarse?"	Hoarseness alters vocal cord vibrations and may affect the system's ability to perceive emotions. While the system may face limitations in such cases, robustness can be improved through larger datasets and adaptive model optimization.
"Can I use the system freely? Are there any limitations?"	Our system works with everyday clothing even when the throat is partially covered. It is robust to different orientations within a 1 m range and handles natural movements, making it suitable for both mobile devices (e.g., smartphones) and fixed platforms (e.g., laptops, in-vehicle systems).

Finer-grained Emotion Perception. *FeelWave* currently recognizes 6 common emotions: happy, calm, angry, tense, sad, and bored. It cannot yet distinguish finer-grained emotional states (e.g., frustrated vs. angry, excited vs. happy) or handle challenging vocal conditions (e.g., hoarseness), primarily due to limited training data. To address this, leveraging the consistency between mmWave vocal vibrations and voice in source features, we plan to use generative methods with large-scale, emotion-rich audio corpora to expand the scarce mmWave dataset and enhance its emotional granularity and generalizability. Notably, we have demonstrated the effectiveness of mmWave vocal signals for emotion recognition, positioning them as a noise-resilient acoustic alternative to conventional audio and a scalable direction for advancing emotion-aware agents. Building on this, we will explore transforming the noise-immune mmWave acoustic features into emotional descriptions and integrating them with semantics to enable robust, fine-grained emotion sensing.

More Users. *FeelWave* currently employs a SISO mmWave radar to minimize power consumption. However, this configuration lacks angular resolution, making it difficult to distinguish multiple users at the same range bin. In future work, adopting multi-input multi-output (MIMO) radars could enable angle-based separation, thereby enabling simultaneous emotion analysis across multiple users and supporting shared use.

Broader Applicability. *FeelWave* leverages mmWave radar to sense vocal cord vibrations for noise-robust emotion recognition. Its performance may degrade at longer distances, when users face away, or when obstacles such as walls are present. To broaden applicability, in low-noise scenarios that require more flexible interaction, such as smart home control or voice assistants in museums, a pre-trained classifier can be used to estimate the vocal content in mmWave signals. When vocal content is low, user emotions can instead be inferred with an audio-based model. Replacing the mmWave module with an audio model preserves the pipeline while extending applicability to more diverse scenarios.

7 CONCLUSION

In this work, we present *FeelWave*, an emotion-aware voice interaction system that enables robust mmWave-based emotion sensing and emotion-driven LLM response generation. By integrating motion-robust signal extraction, cross-domain transfer learning, and emotion-aware query optimization, the system forms a cohesive pipeline for efficient, emotion-driven voice interaction with LLMs. Extensive evaluations show that *FeelWave* achieves 92.3% emotion recognition accuracy and remains robust in real-world noisy environments (86.1% vs. 23.2% for audio-based models). In voice interaction studies, 74.3% of users prefer *FeelWave*, reporting significantly higher satisfaction than a baseline without emotion sensing (4.37 ± 1.23 vs. 3.22 ± 1.03). A SUS score of 88.3 further confirms *FeelWave*'s high usability in the real world. We hope this work will provide new insights into human-AI interaction, advancing systems that understand not only what users say but also how they feel, thereby fostering more empathetic AI-driven assistants.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62332016).

References

[1] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning (ICML)*. 265–279.

[2] Foteini Agrafioti, Dimitris Hatzinakos, and Adam K Anderson. 2011. ECG pattern analysis for emotion detection. *IEEE Transactions on affective computing* 3, 1 (2011), 102–115.

[3] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications* 17 (2023), 200171.

[4] Icek Ajzen and Thomas J Madden. 1986. Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of experimental social psychology* 22, 5 (1986), 453–474.

[5] Paavo Alku. 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication* 11, 2-3 (1992), 109–118.

[6] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 34–44.

[7] Anthropic. 2024. Claude 3.5 Haiku. [Online]. https://www.anthropic.com/claude/haiku.

[8] Apple. 2024. Introducing Apple Intelligence, the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac. [Online]. https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/.

[9] Li Auto. 2024. Li Auto Introduces Fully Self-Developed MindGPT. [Online]. https://genaigazette.com/li-auto-introduces-fully-self-developed-mind-gpt.

[10] Azcom CabinGuard. 2025. A Vehicle In-cabin Monitoring Solution. [Online]. https://www.azcomtech.com/markets/mmwave-radar-sensors/automotive.

[11] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[12] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[13] Felix Burkhardt, A. Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Interspeech 2005*. 1517–1520.

[14] Zhaoxin Chang, Fusang Zhang, Jie Xiong, Weiyan Chen, and Daqing Zhang. 2024. MSense: Boosting Wireless Sensing Capability Under Motion Interference. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking(MobiCom)*. 108–123.

[15] Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2025. Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance. *arXiv preprint arXiv:2510.10444* (2025).

[16] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. From gap to synergy: Enhancing contextual understanding through human-machine collaboration in personalized systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1–15.

[17] Youjun Chen, Xurong Xie, Haoning Xu, Mengzhe Geng, Guinan Li, Chengxi Deng, Huimeng Wang, Shujie Hu, and Xunying Liu. 2025. Towards LLM-Empowered Fine-Grained Speech Descriptors for Explainable Emotion Recognition. In *Interspeech 2025*. 4633–4637.

[18] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 110805–110853.

[19] Pedro Corrêa, João Lima, Victor Moreno, and Paula Dornhofer Paro Costa. 2025. Evaluating Emotion Recognition in Spoken Language Models on Emotionally Incongruent Speech. *arXiv preprint arXiv:2510.25054* (2025).

[20] Livija Cveticanin. 2012. Review on mathematical and mechanical models of the vocal cord. *Journal of Applied Mathematics* 2012, 1 (2012), 928591.

[21] Google DeepMind. 2024. Gemma: Lightweight Open Models for Responsible AI. [Online]. https://ai.google.dev/gemma.

[22] Google DeepMind. 2025. Gemini 2.5 Pro. [Online]. https://blog.google/products/gemini/gemini-2-5-pro-latest-preview/.

[23] DeepSeek. 2024. DeepSeek V3. [Online]. https://github.com/deepseek-ai/DeepSeek-V3.

[24] Han Ding, Yizhan Wang, Hao Li, Cui Zhao, Ge Wang, Wei Xi, and Jizhong Zhao. 2022. Ultraspeech: Speech enhancement by interaction between ultrasound and speech. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 3 (2022), 1–25.

[25] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana. 2014. Glottal source processing: From analysis to applications. *Computer Speech & Language* 28, 5 (2014), 1117–1138.

[26] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition* 44, 3 (2011), 572–587.

[27] Long Fan, Lei Xie, Xinran Lu, Yi Li, Chuyu Wang, and Sanglu Lu. 2023. mmmic: Multi-modal speech recognition based on mmwave radar. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.

[28] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 6491–6501.

[29] Cathy Mengying Fang, Phoebe Chua, Samantha WT Chan, Joanne Leong, Andria Bao, and Pattie Maes. 2025. Leveraging AI-Generated Emotional Self-Voice to Nudge People towards their Ideal Selves. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–20.

[30] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. 2024. Physiollm: Supporting personalized health insights with wearables and large language models. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 1–8.

[31] Federal Communications Commission (FCC). 2021. RF Safety Guidelines. [Online]. https://www.fcc.gov/general/radio-frequency-safety-0.

[32] Guy William Fincham, Clara Strauss, Jesus Montero-Marin, and Kate Cavanagh. 2023. Effect of breathwork on stress and mental health: A meta-analysis of randomised-controlled trials. *Scientific Reports* 13, 1 (2023), 432.

[33] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research (JMLR)* 17, 59 (2016), 1–35.

[34] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 776–780.

[35] Swapna Mol George and P Muhamed Ilyas. 2024. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing* 568 (2024), 127015.

[36] Google. 2017. Webrtc-vad. [Online]. https://webrtc.org/.

[37] Google. 2021. ok-google.io. [Online]. https://ok-google.io.

[38] Google. 2025. Speech-to-Text. [Online]. https://cloud.google.com/speech-to-text.

[39] Feiyu Han, Panlong Yang, You Zuo, Fei Shang, Fenglei Xu, and Xiang-Yang Li. 2024. Earspeech: Exploring in-ear occlusion effect on earphones for data-efficient airborne speech enhancement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 8, 3 (2024), 1–30.

[40] Feiyu Han, You Zuo, Weiwei Jiang, Dawei Yan, Yuxin Zhao, Panlong Yang, and Yubo Yan. 2025. EAROE: Enabling Body-Channel Voice Interaction Interface on Earphones via Occlusion Effect. *IEEE Internet of Things Journal (IoTJ)* (2025).

[41] Hardso. 2025. AIBI Pocket. [Online]. https://www.hardso.com/product/2fea3b01-c2a1-4050-8e7e-f9d62f88de88.

[42] Hardso. 2025. PieX Pendant. [Online]. https://www.hardso.com/product/ea05912b-a7fc-4bd5-aee4-f43f8ac4d200.

[43] Chenming He, Chengzhen Meng, Chunwang He, Xiaoran Fan, Beibei Wang, Yubo Yan, and Yanyong Zhang. 2024. See Through Vehicles: Fully Occluded Vehicle Detection with Millimeter Wave Radar. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 740–754.

[44] Chenming He, Rui Xia, Chengzhen Meng, Xiaoran Fan, Dequan Wang, Haojie Ren, Jianmin Ji, and Yanyong Zhang. 2025. Ghost Points Matter: Far-Range Vehicle Detection with a Single mmWave Radar in Tunnel. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking (MobiCom)*. 650–666.

[45] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 131–135.

[46] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[47] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*. 1314–1324.

[48] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2022. The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses. *IEEE Transactions on Affective Computing* 14, 1 (2022), 17–30.

[49] Infineon. 2020. Radar Development Kit. [Online]. https://www.infineon.com/cms/en/design-support/tools/sdk/radar-development-kit/.

[50] Infineon. 2023. bgt60tr13c. [Online]. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60tr13c.

[51] Infineon. 2023. bgt60utr11aip. [Online]. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60utr11aip/.

[52] Texas Instruments. 2020. IWR6843 intelligent mmWave sensor standard antenna plug-in module. [Online]. https://www.ti.com/tool/IWR6843ISK.

[53] Texas Instruments. 2024. IWR6843. [Online]. https://www.ti.com/product/IWR6843#order-quality.

[54] International Commission on Non-Ionizing Radiation Protection (ICNIRP). 2020. Guidelines for Limiting Exposure to Electromagnetic Fields (100 kHz to 300 GHz). [Online]. https://www.icnirp.org/cms/upload/publications.

[55] Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).

[56] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. 2011. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th international colloquium on signal processing and its applications*. IEEE, 410–415.

[57] Samuel Kakuba and Dong Seog Han. 2025. Addressing data scarcity in speech emotion recognition: A comprehensive review. *ICT Express* 11, 1 (2025), 110–123.

[58] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. Emotions on the go: Mobile emotion assessment in real-time using facial expressions. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces (AVI)*. 1–9.

[59] Adi Lausen and Kurt Hammerschmidt. 2020. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–17.

[60] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760* (2023).

[61] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).

[62] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).

[63] Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836* (2024).

[64] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. [n. d.]. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning (ICML)*.

[65] Yu-Xiang Lin, Chih-Kai Yang, Wei-Chih Chen, Chen-An Li, Chien-yu Huang, Xuanjun Chen, and Hung-yi Lee. 2025. A preliminary exploration with gpt-4o voice mode. *arXiv preprint arXiv:2502.09940* (2025).

[66] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. 97–110.

[67] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. 2023. WavoID: Robust and secure multi-modal user identification via mmWave-voice mechanism. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1–15.

[68] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[69] Andrey Lovakov and Elena R Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology* 51, 3 (2021), 485–504.

[70] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*. 1–12.

[71] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics (ACL)*. 15747–15760.

[72] Brian McFee, Justin Salamon, and Juan Pablo Bello. 2018. Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 26, 11 (2018), 2180–2193.

[73] Chengzhen Meng, Yifan Duan, Chenming He, Dequan Wang, Xiaoran Fan, and Yanyong Zhang. 2024. mmPlace: Robust Place Recognition With Intermediate Frequency Signal of Low-Cost Single-Chip Millimeter Wave Radar. *IEEE Robotics*

and Automation Letters (RAL) 9, 6 (2024), 4878–4885.
[74] Chengzhen Meng, Chenming He, Dequan Wang, Yuxuan Xiao, Lingyu Wang, Xiaoran Fan, Lu Zhang, and Yanyong Zhang. 2025. Gr-fall: A fall detection system with gait recognition for indoor environments using siso mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 9, 3 (2025), 1–26.
[75] Meta AI. 2024. Llama 3.1. [Online]. https://ai.meta.com/blog/meta-llama-3-1/.
[76] ICLEGEND MICRO. 2024. 24GHz mmWave Sensor SoC. [Online]. https://www.iclegend.com/zh-hans/product/category/Sensor.
[77] NOVELIC. 2025. A Whole-Cabin Solution for Vehicle Safety and Comfort. [Online]. https://www.novelic.com/acam-automotive-in-cabin-monitoring-radar/.
[78] OpenAI. 2024. GPT-4o. [Online]. https://openai.com/index/hello-gpt-4o/.
[79] OpenAI. 2024. GPT-4o Audio Preview. [Online]. https://platform.openai.com/docs/models/gpt-4o-audio-preview.
[80] OpenAI. 2024. Introducing Our Next-Generation Audio Models. [Online]. https://openai.com/index/introducing-our-next-generation-audio-models/.
[81] OpenAI. 2025. GPT-4o-mini. [Online]. https://platform.openai.com/docs/models/gpt-4o-mini.
[82] OpenAI. 2025. GPT-4o-mini-TTS. [Online]. https://platform.openai.com/docs/models/gpt-4o-mini-tts.
[83] OpenAI. 2025. GPT-4o-transcribe. [Online]. https://platform.openai.com/docs/models/gpt-4o-transcribe.
[84] OpenAI. 2025. OpenAI Security and Privacy. [Online]. https://openai.com/security-and-privacy/?utm_source=chatgpt.com.
[85] Alan Oppenheim and Ronald Schafer. 2013. *Discrete-Time Signal Processing* (3 ed.). Pearson Deutschland.
[86] Geoffrey Ottoy, Bart Thoen, and Lieven De Strycker. 2016. A low-power MEMS microphone array for wireless acoustic sensors. In *IEEE Sensors Applications Symposium (SAS)*. 1–6.
[87] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. 2023. Radio SES: mmWave-Based Audioradio Speech Enhancement and Separation System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 31 (2023), 1333–1347.
[88] Ashutosh Pandey and DeLiang Wang. 2019. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 27, 7 (2019), 1179–1188.
[89] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–20.
[90] Rosalind W Picard. 2000. *Affective computing*. MIT press.
[91] Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, and James R. Glass. 2025. Quantifying Generalization Complexity for Large Language Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
[92] Maija Reblin and Bert N Uchino. 2008. Social and emotional support and its implication for health. *Current opinion in psychiatry* 21, 2 (2008), 201–205.
[93] Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, and Longfei Shangguan. 2025. Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*. 254–267.
[94] Google Research. 2020. Soli Radar-based Perception and Interaction in Pixel 4. [Online]. https://research.google/blog/soli-radar-based-perception-and-interaction-in-pixel-4/.
[95] RESEMBLE.AI. 2023. Introducing Resemble Enhance: Open Source Speech Super Resolution AI Model. [Online]. https://www.resemble.ai/introducing-resemble-enhance/.
[96] Mark A. Richards. 2005. *Fundamentals of Radar Signal Processing*. McGraw-Hill.
[97] Chantel Ritter and Tara Vongpaisal. 2018. Multimodal and spectral degradation effects on speech and emotion recognition in adult listeners. *Trends in Hearing* 22 (2018).
[98] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 5986–6004.
[99] Samsung. 2021. mmWave 5G: Past, Present and Future. [Online]. https://www.samsung.com/global/business/networks/insights/blog/0218-mmwave-5g-past-present-and-future/.
[100] Katja Schlegel, Nils R Sommer, and Marcello Mortillaro. 2025. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology* 3, 1 (2025), 80.
[101] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.
[102] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards Learning a Universal Non-Semantic Representation of Speech. In *Interspeech 2020*. 140–144.
[103] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. 37–76.
[104] Maya Tamir. 2016. Why do people regulate their emotions? A taxonomy of motives in emotion regulation. *Personality and social psychology review* 20, 3 (2016), 199–222.
[105] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[106] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R Cowan, and Heinrich Hussmann. 2021. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems (CHI)*. 1–15.
[107] Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024. BLSP-Emo: Towards Empathetic Large Speech-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 19186–19199.
[108] Dequan Wang, Chenming He, Lingyu Wang, Chengzhen Meng, Xiaoran Fan, and Yanyong Zhang. 2026. FlowGait: Enabling Robust Long-Term Gait Recognition Across Real-World Covariates with mmWave Radar. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM New York, NY, USA, 22 pages.
[109] Dequan Wang, Xinran Zhang, Kai Wang, Lingyu Wang, Xiaoran Fan, and Yanyong Zhang. 2024. Rdgait: A mmwave based gait user recognition system for complex indoor environments using single-chip radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 8, 3 (2024), 1–31.
[110] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology* 17 (2023).
[111] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
[112] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[113] Xin-Cheng Wen, JiaXin Ye, Yan Luo, Yong Xu, Xuan-Ze Wang, Chang-Li Wu, and Kun-Hong Liu. 2022. CTL-MTNet: A Novel CapsNet and Transfer Learning-Based Mixed Task Net for Single-Corpus and Cross-Corpus Speech Emotion Recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*. 2305–2311.
[114] Wikipedia. 2024. Voice frequency. [Online]. https://en.wikipedia.org/wiki/Voice_frequency.
[115] Wikipedia. 2025. Tukey window (cosine-tapered window). [Online]. https://en.wikipedia.org/wiki/Window_function.
[116] Mingyang Wu and DeLiang Wang. 2006. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)* 14, 3 (2006), 774–784.
[117] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2025. Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances. In *Findings of the Association for Computational Linguistics (NAACL)*. 2202–2218.
[118] Zhongzhe Xiao, Ying Chen, and Zhi Tao. 2018. Contribution of glottal waveform in speech emotion: A comparative pairwise investigation. In *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 185–190.
[119] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. 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)*. 14–26.
[120] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 269–282.
[121] Huanpu Yin, Shuhui Yu, Yingshuo Zhang, Anfu Zhou, Xin Wang, Liang Liu, Huadong Ma, Jianhua Liu, and Ning Yang. 2022. Let iot know you better: User identification and emotion recognition through millimeter-wave sensing. *IEEE Internet of Things Journal (IoTJ)* 10, 2 (2022), 1149–1161.
[122] JTFLM Zhang and Huibin Jia. 2008. Design of speech corpus for mandarin text to speech. In *The blizzard challenge 2008 workshop*.

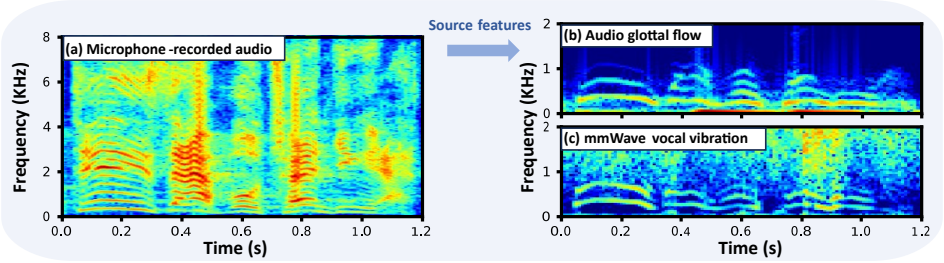

Figure 27: Acoustic feature analysis. (a) Microphone-recorded audio presents both glottal source and vocal tract features with clear fundamental and harmonic components; (b)-(c) The mmWave-captured vocal signal and audio-derived glottal flow exhibit strong similarity, demonstrating the cross-modal consistency of source features.

[123] Xi Zhang, Yu Zhang, Zhenguo Shi, and Tao Gu. 2023. mmfer: Millimetre-wave radar based facial expression recognition for multimedia iot applications. In *Proceedings of the 29th annual international conference on mobile computing and networking (MobiCom)*. 1–15.

[124] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd annual international conference on mobile computing and networking*. 95–108.

[125] Running Zhao, Jiangtao Yu, Hang Zhao, and Edith CH Ngai. 2023. Radio2Text: Streaming Speech Recognition Using mmWave Radio Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 7, 3 (2023), 1–28.

[126] Hengshun Zhou, Jun Du, Yan-Hui Tu, and Chin-Hui Lee. 2020. Using Speech Enhancement Preprocessing for Speech Emotion Recognition in Realistic Noisy Conditions. In *Interspeech 2020*. 4098–4102.

A Appendix

A.1 Preliminary

A.1.1 Principles of mmWave Radar Sensing. mmWave technology is gaining increasing attention across diverse applications [43, 44, 73, 74, 108, 109], enabled by its high-precision range and velocity measurements and its growing affordability. In FMCW radar systems, the signal propagation delay induces a characteristic beat frequency between the transmitted and received signals, resulting in intermediate frequency (IF) signals expressed as:

$$IF(t) = A_t A_r e^{j2\pi(\rho t + f_0)\tau}, \tag{6}$$

where A_t and A_r are the amplitudes of the transmitted and received signals, ρ is the chirp rate, f_0 is the starting frequency, and τ denotes the propagation delay. Thus, the frequency of the IF signal can be expressed as $f_{IF} = 2\pi\rho\tau$.

Range Estimation. A frequency shift arises from the time of flight between transmitted and received signals. The IF signal frequency (f_{IF}) is proportional to the radar-subject distance: $d = \frac{f_{IF}c}{2\rho}$, where c is the speed of light.

Velocity Estimation. For a moving subject, the phase of the IF signal changes with the distance between radar and the subject:

$$\varphi \approx 2\pi f_0 \tau = \frac{4\pi d}{\lambda}, \tag{7}$$

where λ is the wavelength. Therefore, the phase difference between consecutive chirps is proportional to the subject's Doppler velocity as $v = \frac{\lambda\Delta\varphi}{4\pi T_c}$, where T_c is the time interval between multiple chirps.

A.1.2 Vibration-Based Voice Production. Voice serves as a primary channel for emotional expression, conveying different emotions through variations in pitch, volume, and rhythm. During phonation, air from the lungs passes through the vocal cords and is shaped by the vocal tract to produce speech. As shown in Fig. 2, vocal cord vibrations serve as the primary source, playing a key role in defining speech characteristics such as the fundamental frequency and partial harmonics. Their dynamic relationship [20, 67] can be modeled as:

$$\begin{aligned} m\ddot{u}(t) + r\dot{u}(t) + ku(t) &= F_0 e^{j(2\pi f_F t + \theta_F)}, \\ s(t) &= \mathcal{T}(\dot{u}(t)), \end{aligned} \tag{8}$$

where $u(t)$ is the vocal cord displacement, and m, r, and k are parameters governed by vocal cord physiology. The force $F_0 e^{j(2\pi f_F t + \theta_F)}$ varies with the degree of vocal fold tightness. The transfer function $\mathcal{T}(\cdot)$ maps the velocity $\dot{u}(t)$ to the radiated speech signal $s(t)$. As the core of voice production, vocal vibrations encode emotional cues such as pitch, intensity, and rhythm, making them well-suited for emotion recognition.

A.1.3 Vocal Vibration Signal Sensing. Furthermore, due to the millimeter-scale wavelength of mmWave, the phase variations enable detection of micro-movements. Accordingly, based on Equation 7, the phase difference ($\Delta\varphi_v$) resulting from vocal vibration can be expressed as:

$$\Delta\varphi_v = \frac{4\pi\delta}{\lambda}, \tag{9}$$

where δ is the displacement due to vocal cord vibration. Since $\Delta\varphi_v$ is linearly proportional to δ, it represents the vocal vibration signal and contains the fundamental frequency as well as partial harmonics of the voice, as shown in Fig. 2.

Cross-Modal Acoustic Analysis. Source features, as the core components of voice production, encode rich emotional cues [118]. They describe the vocal cord vibration patterns and their excitation signals, which are likewise captured by mmWave phase differences. To examine the relationship between mmWave and microphone-recorded audio modalities, we focus on emotion-relevant source features and analyze their correlation. We apply iterative adaptive inverse filtering (IAIF) [5] to both signals to extract the *audio-derived glottal flow* and compute *glottal-source linear predictive coefficients (LPCs)* for both modalities [25]. As shown in Fig. 27, the mmWave-captured vocal signal and audio glottal flow share the fundamental frequency and partial harmonics, showing strong consistency. Furthermore, the glottal-source LPCs from both modalities achieve an average cosine similarity of 0.81 across ten female and ten male participants, highlighting the value of mmWave-extracted vocal features for emotion recognition.

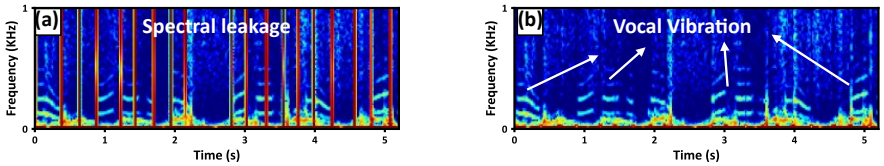

Figure 28: Signal smoothing process. (a) Spectral leakage caused by phase discontinuities at window edges; (b) phase difference signal after smoothing.

A.2 mmWave Signal Processing Pipeline

A.2.1 Details of Vocal-Intensive Bin Selection. According to the range resolution $d_{\text{res}} = \frac{c}{2B}$, a radar bandwidth B of 3.6 GHz yields a resolution of 0.042 m. Empirically, a neighborhood size of $N = 7$ adjacent range bins covers roughly 0.3 m, accommodating the typical extent of user motion in daily use. As described in Section 3.1.1, we first locate the range bin $r_{\max}$ with the highest energy and define its IQ neighborhood $\mathcal{N}_{\text{neigh}}$. The following provides the full implementation details of vocal-intensive bin selection, including the vocal energy estimator and the smoothing process.

Vocal Energy Estimator. The vocal vibration frequencies captured by radar typically remain below 1000 Hz. Therefore, within a time window W set to 0.2 s, each phase difference signal computed from $\mathcal{N}_{\text{neigh}}$ is filtered into three sub-bands: 80-250 Hz, 250-500 Hz, and 500-1000 Hz. These filtered signals are denoted as $\{\Delta\phi_{s_m}^{(1)}, \Delta\phi_{s_m}^{(2)}, \Delta\phi_{s_m}^{(3)}\}$, where $m = 1, 2, \ldots, N$. For each sub-band, we compute its log energy and the total energy of the full-band signal to form a 4D vector $\mathbf{x}_m = [x_{m,1}, x_{m,2}, x_{m,3}, E_m]^T$, where $x_{m,i} = \log\left(\sum_{t\in W}\left|\Delta\phi_{s_m}^{(i)}(t)\right|^2\right), i = 1, 2, 3$ and $E_m = \log\left(\sum_{t\in W}\left|\Delta\phi_{s_m}(t)\right|^2\right)$. To model the distribution of vocal features in mmWave phase difference signals, we adopt a single Gaussian model, defined as $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$, where $\mathbf{x} \in \mathbb{R}^4$ is the feature vector, $\boldsymbol{\mu} \in \mathbb{R}^4$ is the mean vector, and $\Sigma \in \mathbb{R}^{4\times4}$ is the covariance matrix. To initialize the model, we collect mmWave vocal vibration segments and extract feature vectors $\{\mathbf{x}_k\}_{k=1}^K$. The parameters are estimated via maximum likelihood as $\boldsymbol{\mu} = \frac{1}{K}\sum_{k=1}^K \mathbf{x}_k$ and $\Sigma = \frac{1}{K}\sum_{k=1}^K(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T$. The IQ signal with the most vocal content is selected as:

$$S_{\text{opt}} = \text{argmax}_{s_m\in\mathcal{N}_{\text{neigh}}}\, p(\mathbf{x}_m). \tag{10}$$

To maintain model adaptability under long-term variations in speech characteristics, we adopt a slow update strategy. When $p(\mathbf{x}_{\text{opt}}) \geq 0.5$, the model parameters are updated using a small learning rate $\beta \in [10^{-3}, 10^{-2}]$ as $\boldsymbol{\mu}_{\text{new}} = (1-\beta)\boldsymbol{\mu}_{\text{old}} + \beta\mathbf{x}_{\text{opt}}$ and $\Sigma_{\text{new}} = (1-\beta)\Sigma_{\text{old}} + \beta(\mathbf{x}_{\text{opt}} - \boldsymbol{\mu}_{\text{new}})(\mathbf{x}_{\text{opt}} - \boldsymbol{\mu}_{\text{new}})^T$.

Signal Smoothing. However, phase discontinuities at the edges of consecutive time windows can still introduce spectral leakage, as shown in Fig. 28(a). To mitigate this, we apply Tukey windowing to create smooth fade-in and fade-out regions at both ends of the signal. The signal S_{opt} is of length N, with left and right smoothing regions of lengths N_{in} and N_{out}, respectively. The smoothing window is defined as:

$$w[n] = \begin{cases} w_{\text{in}}[n], & 1 \leq n \leq N_{\text{in}} \\ 1, & N_{\text{in}} < n \leq N - N_{\text{out}} \\ w_{\text{out}}[n-(N-N_{\text{out}})], & N - N_{\text{out}} < n \leq N \end{cases}. \tag{11}$$

The fade-in and fade-out windows are extracted from a full Tukey window [115] as follows:

$$\begin{aligned} w_{\text{in}}[n] &= \text{Tukey}(2N_{\text{in}}, \alpha)[n], \quad 1 \leq n \leq N_{\text{in}}, \\ w_{\text{out}}[n] &= \text{Tukey}(2N_{\text{out}}, \alpha)[n + N_{\text{out}}], \quad 1 \leq n \leq N_{\text{out}}. \end{aligned} \tag{12}$$

The smoothed IQ signal is given by $\tilde{S}_{\text{opt}} = S_{\text{opt}} \cdot w$. As shown in Fig. 28(b), after signal smoothing, spectral leakage in the phase difference of the signal is effectively suppressed.

A.2.2 Preliminary Study of Body Motion Demodulation. The phase difference $\Delta\varphi_v$ captures vocal vibrations but is also susceptible to motion-induced interference. As shown in Fig. 5(a) and (b), motion-induced components overlap with vocal vibration frequencies, causing distortion. The relationship between phase difference frequency f and subject velocity v is expressed as:

$$v = \frac{\lambda\Delta\varphi}{4\pi T_c} \Rightarrow f = \frac{2\Delta v}{\lambda}. \tag{13}$$

The corresponding acceleration is $a = \frac{\lambda}{2T}f$. With a chirp interval $T = 100\ \mu$s and wavelength $\lambda \approx 5$ mm, motion-induced frequencies above vocal fundamentals (e.g., 90Hz [114]) require acceleration above 2250 m/s², *a physiologically unattainable level*, which contradicts our experimental observations. To investigate the cause of the overlapping distortion, we conducted experiments as follows.

Simulation of Motion Interference. To simulate motion-induced interference in mmWave radar sensing, we design a hybrid motion profile comprising: (1) constant linear motion, (2) uniformly decelerated motion, and (3) a stationary state. These motion patterns are used to generate the corresponding Doppler signal. We adopt a radar configuration with sampling frequency $f_s = 10$ kHz, carrier frequency $f_0 = 60$ GHz, and speed of light $c = 3\times10^8$ m/s. The time index is $t[n] = \frac{n}{f_s}, n = 0, 1, \ldots, N-1$. The velocity sequence $v[n]$ is defined over three segments: uniform (N_1), uniform deceleration (N_2), and stationary ($N_3 = N - N_1 - N_2$) . The motion profile is:

$$v[n] = \begin{cases} 1.0, & 0 \leq n < N_1 \\ 1.0 - \frac{n-N_1}{N_2}, & N_1 \leq n < N_1 + N_2 \\ 0, & N_1 + N_2 \leq n < N \end{cases}. \tag{14}$$

The corresponding Doppler frequency is $f_d[n] = \frac{2v[n]}{c}f_0$. By integrating the frequency over time, the motion-induced IQ signal is synthesized via phase accumulation:

$$\begin{aligned} \phi_d[n] &= \phi[n-1] + 2\pi f_D[n-1]\cdot\Delta t, \quad \Delta t = \frac{1}{f_s}, \\ s_d[n] &= \cos(\phi[n]) + j\sin(\phi[n]) = e^{j\phi[n]}. \end{aligned} \tag{15}$$

We then modulate the recorded static IQ signal $\mathbf{s}_{\text{raw}}[n]$ with motion and a radar DC component ($DC = 600$), yielding $\mathbf{s}_{\text{mod}}[n] = \mathbf{s}_{\text{raw}}[n]\cdot s_d[n] + DC$. Phase dynamics are extracted as: (1) $\tilde{\theta}[n] =$

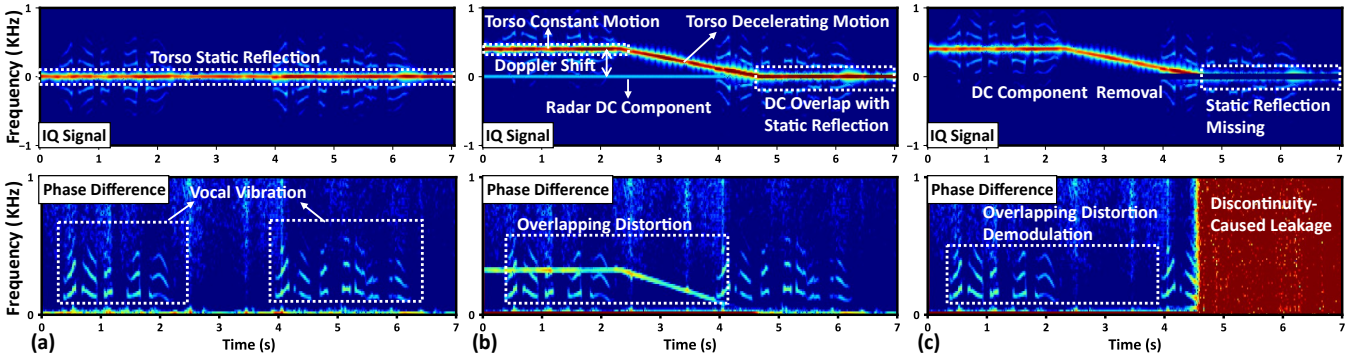

Figure 29: Motion-induced interference simulation on mmWave signals. (a) Vocal vibrations under static conditions; (b) Interference from hybrid motion, causing Doppler shifts in IQ signals and overlapping distortion in phase difference due to combined motion and the DC component; (c) Removing the DC component mitigates overlapping distortion, but simultaneously removing static reflections introduces spectral leakage.

unwrap ($\angle \mathbf{s}_{\text{mod}}[n]$), and (2) $\Delta\theta[n] = \tilde{\theta}[n+1] - \tilde{\theta}[n]$, where $\tilde{\theta}[n]$ denotes the unwrapped instantaneous phase, and $\Delta\theta[n]$ is the phase difference sequence.

Analysis. We record vocal vibrations under static conditions as a reference (Fig. 29(a)). When motion is introduced, the IQ spectrogram in Fig. 29(b) shows that vocal frequencies in $\mathbf{s}_{\text{mod}}$ are modulated by motion-induced shifts. By analyzing the spectrogram of the corresponding phase difference signal ($\Delta\theta$), we observe motion-induced overlapping distortion, which violates the theoretical relationship in Equation 13. Removing the zero-frequency DC component from $\mathbf{s}_{\text{mod}}$ eliminates this distortion in $\Delta\theta$ (Fig. 29(c)). However, since the DC component overlaps with subsequent static reflections, direct removal introduces discontinuities and spectral leakage. These results indicate that DC-motion coupling generates pseudo-frequencies that cause overlapping distortion. Building on this, we propose a motion demodulation algorithm that selectively removes the DC component during movement to eliminate overlap-induced distortion.

A.2.3 Details of the Conditional Smoothed Band-stop Filter. The filter's frequency response is designed as:

$$H_{\text{bs}}(f) = \begin{cases} 0, & |f| \le \Delta \\ \frac{1}{2}\left[1 - \cos\left(\pi \cdot \frac{|f|-\Delta}{\delta}\right)\right], & \Delta < |f| \le \Delta + \delta \\ 1, & |f| > \Delta + \delta \end{cases}, \quad (16)$$

where the stop band is defined as $|f| \le \Delta$, and δ denotes the transition band to mitigate spectral leakage from sharp frequency cutoffs.

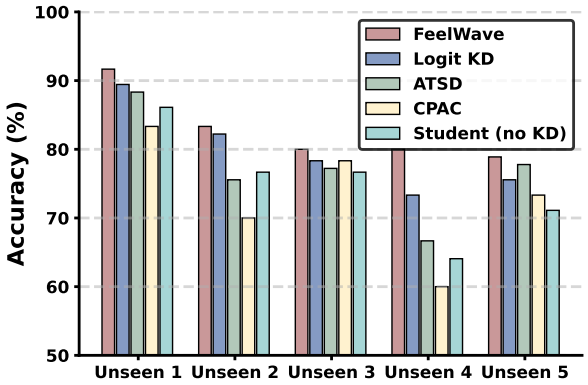

Figure 30: Generalization for unseen users.

A.3 User Queries

To complement the design in Section 4.1.1, Table 5 summarizes the three query types with concise descriptions and representative examples (adapted from ok-google.io [37]).

A.4 Breakdown Statistical Results for Human and LLM Participants

This part reports the per-group statistical analyses for human participants and LLM participants. We report Wilcoxon signed-rank tests and Cohen's d effect sizes across all evaluation metrics and all task types (T1, T2, T3). The detailed results are provided in Table 6 and complement the aggregated findings described in Section 4.2.1. Emotional cues generally produce noticeable improvements for

Table 5: Categories of user queries with descriptions and examples.

Query Type	Description	Examples
Information retrieval	Request factual or context-aware information (e.g., flight status, weather, general knowledge).	*"When will AA125 land?", "Do I need an umbrella for tomorrow?"*
Procedural guidance	Seek step-by-step instructions or navigation to accomplish a task.	*"How do I make an Old Fashioned cocktail?", "How do I get to the supermarket by walking?"*
Task execution	Directive queries where the user expects the system to perform an action or trigger a service.	*"Show me the appointments for tomorrow.", "Play some music."*

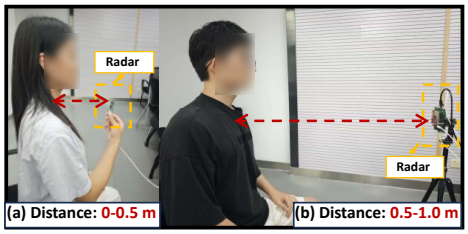

Figure 31: Experimental setup for distance robustness. (a) mobile-device scenarios (e.g., smartphones); (b) fixed-device scenarios (e.g., laptops, in-vehicle systems).

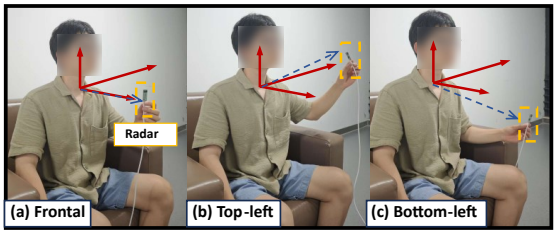

Figure 32: Experimental setup for orientation robustness. (a)-(c) cover an azimuth of up to approximately 60° and elevations of up to +15° and -45°, with a radar FOV of around 40°.

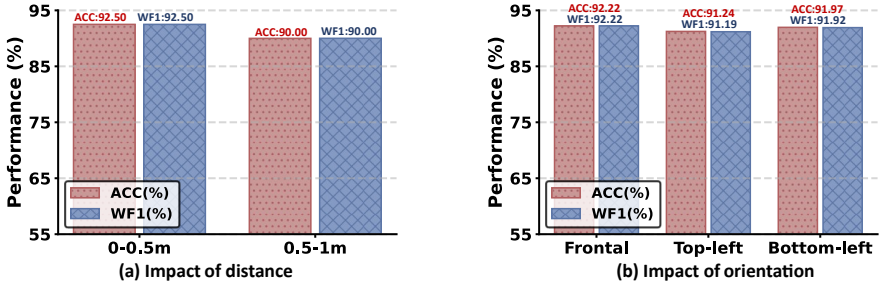

Figure 33: Impact of distance and orientation.

both participant groups, with the strongest effects observed in procedural guidance and task execution. Minor variations in clarity, particularly for human participants in T1, are also included.

A.5 Robustness Evaluation of *FeelWave*'s Emotion Sensing

To evaluate *FeelWave*'s performance across diverse scenarios, we conduct robustness experiments involving unseen speakers, variations in the distance and user-radar orientation, user movements, and everyday clothing occlusion.

A.5.1 Generalization for Unknown Users. We collect mmWave data from 5 unseen volunteers to evaluate generalization performance. Differences in vocal emotion expression and physiological traits among volunteers may pose challenges to the generalization of emotion recognition models based on vocal vibrations. As shown in Fig. 30, *FeelWave* achieves an average emotion recognition ACC of 82.78%, with a maximum of 91.67%. Compared to known users, its performance decreases due to the limited scale of the mmWave dataset, but it still demonstrates effective generalization capability. Furthermore, *FeelWave* achieves superior generalization, exceeding the performance of non-distillation methods by approximately 9 percentage points. These results indicate that the cross-transfer approach can leverage large-scale audio data to guide feature learning with small-scale mmWave data, thereby achieving superior generalization.

A.5.2 Impact of Distance and Orientation. We systematically investigate the impact of the user-device distance, as well as the radar's orientation relative to the participant, on *FeelWave*. The experimental setups for distance and orientation are shown in Fig. 31 and Fig. 32, respectively. For distance robustness testing, we define two ranges: 0-0.5 m and 0.5-1.0 m, corresponding to everyday usage scenarios for mobile devices (e.g., smartphones) and fixed devices (e.g., laptops, in-vehicle systems), respectively. For orientation, we classify the radar's position relative to the user into three ranges: Frontal, Top-left, and Bottom-left, covering an azimuth of up to approximately 60°, elevations of up to +15° and -45°, and a horizontal and vertical FOV of approximately 40°.

As shown in Fig. 33, the emotion perception performance of *FeelWave* slightly decreases with distance but remains above 90%.

Table 6: Cohen's d and statistical significance of human vs. LLM participants across metrics and query types.

	Human participants				LLM participants			
Metric	T1	T2	T3	All	T1	T2	T3	All
Clarity	-0.35*	**1.20 ✓**	**0.65 ✓**	**0.37 ✓**	0.02†	**0.42 ✓**	0.13†	0.20
Emotional Appropriateness	**1.81 ✓**	**2.06 ✓**	**1.79 ✓**	**1.85 ✓**	**2.00 ✓**	**1.70 ✓**	**1.54 ✓**	**1.75 ✓**
Contextual Relevance	**0.89 ✓**	**1.41 ✓**	**1.48 ✓**	**1.18 ✓**	**0.57 ✓**	**0.50 ✓**	**0.61 ✓**	**0.55 ✓**
Interaction Comfort	**1.44 ✓**	**1.61 ✓**	**1.75 ✓**	**1.58 ✓**	**1.37 ✓**	**1.28 ✓**	**1.03 ✓**	**1.23 ✓**
Overall Satisfaction	**1.07 ✓**	**1.60 ✓**	**1.86 ✓**	**1.41 ✓**	**1.79 ✓**	**1.34 ✓**	**1.22 ✓**	**1.44 ✓**

Note: Each cell reports Cohen's d effect size. Superscript † indicates $p \geq 0.05$ (not significant). Values in bold with ✓ mark significant effects with $d \geq 0.36$ (medium or large).

* **For human raters in T1, emotional cues slightly reduce perceived clarity (from 4.36 to 4.14 on average).**

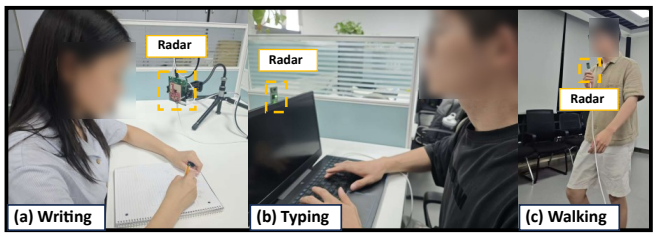

Figure 34: Experimental setup for motion robustness. (a)-(c) include potential movements of the head, torso, and limbs.

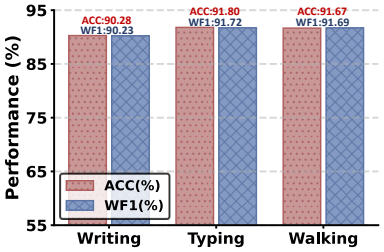

Figure 35: Impact of user movements.

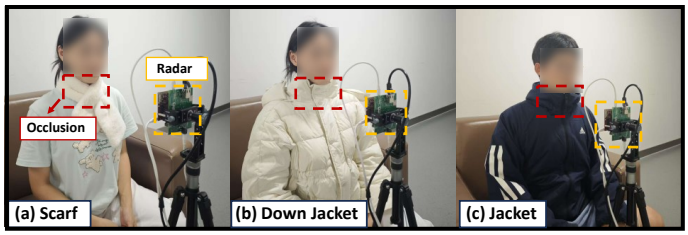

Figure 36: Experimental setup for occlusion robustness. (a)-(c) include occlusion of the vocal cord by everyday clothing of varying thickness.

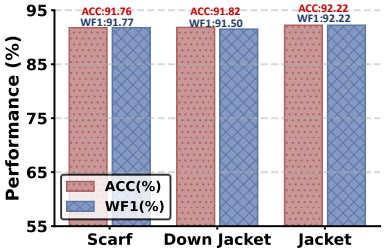

Figure 37: Impact of clothing occlusion.

Furthermore, *FeelWave* demonstrates strong robustness across different orientations, including Frontal, Top-left, and Bottom-left, maintaining an average ACC and WF1 of 91.78%. In extreme cases, such as distances exceeding 1 m or when the vocal cord region is not visible to the radar, the radar cannot capture complete vocal vibration information. However, such cases are rare in everyday use. Therefore, *FeelWave* can be effectively applied in typical mobile-device scenarios, such as voice interaction on smartphones, as well as in fixed-device scenarios, such as in-vehicle voice assistants, where it can robustly perceive emotions under natural usage distances and orientations.

A.5.3 Impact of User Movements. As shown in Fig. 36, we record mmWave data during writing, typing, and walking, where the user exhibits natural movements of the head, torso, and limbs. As shown in Fig. 35, *FeelWave* achieves an average ACC of 91.25% and an average WF1 of 91.21%, demonstrating robustness to natural user movements. A slight performance drop is observed during the writing activity, which is attributed to partial occlusion of the vocal cord region caused by head movements. Overall, the motion-robust vocal signal extraction algorithm enables *FeelWave* to maintain robustness under typical daily movements.

A.5.4 Impact of Clothing Occlusion. To evaluate the performance of *FeelWave* under everyday clothing occlusion, we use a scarf, a down jacket, and a jacket to cover the vocal cord region, as shown in Fig. 36. As illustrated in Fig. 37, mmWave signals can effectively penetrate everyday clothing to capture vocal vibrations, enabling *FeelWave* to achieve an average ACC of 91.93% and an average WF1 of 91.83%. These results indicate that *FeelWave* is robust to everyday clothing occlusion while maintaining effective emotion perception. However, mmWave signals cannot penetrate occlusion caused by body parts or materials such as metal, under which *FeelWave* may fail to operate effectively.

A.6 Interview Details

As described in Section 5.2, the interviews involve 4 university students aged 23-27, an office worker aged 31, a taxi driver aged 45, a professor aged 51, and a retired senior aged 67, offering perspectives that reflect diverse daily routines, emotional needs, and varying familiarity with AI interactions. Among them, the office worker and student participants directly interact with *FeelWave*, while the others learn about the system through a video demonstration. In particular, for the retired participant, the demonstration is accompanied by verbal explanation due to limited prior familiarity with such applications. We summarize their comments as follows:

Student 1: *"I find emotion-aware voice interaction novel and engaging. It could potentially be extended to applications such as personal digital pets or other interactive games."*

Student 2: *"I really enjoy listening to music, and it would be exciting if the system could recommend songs that match my mood."*

Office worker: *"My commute and heavy workload often make me anxious. The system detects my urgency and organizes a clear schedule, which eases my stress."*

Taxi driver: *"In traffic jams or facing reckless drivers, I get irritated. This system senses my emotions, helps me relax, and keeps me focused on driving, reducing impulsive behavior."*

Professor: *"I like that this system uses emotions to uncover deeper needs and actually solve problems, rather than just offering verbal comfort. It would be an interesting extension if it could support multiple users, not just a personal agent, by linking each member's emotions and preferences while accounting for both emotional and individual differences."*

Retired senior: *"I'm not familiar with this kind of interaction, and I do get bored at times. If the system could talk with me like a well-informed friend, sharing recent happenings and keeping me company, I would welcome it."*

These excerpts complement the exploratory analysis by illustrating participants' perspectives on the practical value and future directions of *FeelWave*.