2024 IEEE International Performance, Computing, and Communications Conference (IPCCC) | 979-8-3503-6794-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IPCCC59868.2024.10850088

DASIV: Directional Acoustic Sensing based Intelligent Vehicle Interaction System

Dinghua Zhao †‡&, Juntao Zhou †‡&, Dian Ding†‡*, Yu Lu †‡, Yijie Li†‡, Hang Yang†‡, Yi-Chao Chen†‡, Guangtao Xue†‡
† Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
‡ Shanghai Key Laboratory of Trusted Data Circulation, Governance and Web3
Email:{zhaodinghua, juntaozhou, dingdian94, yulu01, yijieli, yanghang_0505, yichao, gt_xue}@sjtu.edu.cn

***Abstract*—With the increase in motor vehicles, more convenient and accurate interactions are expected while retaining a high standard of safe driving. However, complex and dynamic vehicle environments challenge sensing tasks such as breathing monitor and hand gesture recognition. In this paper, we propose DASIV, which utilizes the highly directional nature of ultrasonic signals to achieve fine-grained directional acoustic sensing in vehicle environments. Due to air nonlinearity, the system enables synchronized directional acoustic communication to transmit information (e.g., navigation) to the driver without affecting other passengers. By optimizing the frequency of the Frequency Modulated Continuous Wave (FMCW) signals, DASIV avoids mutual interference between the sensing and communication signals and achieves breathing detection and hand gesture recognition for the driver. Specifically, the system extracts breathing-induced weak thoracic bullying through the signal phase, captures and analyses breathing patterns using bandpass and Gaussian filters, and develops a breathing model. Then, the system defines 10 interaction hand gestures to meet daily interaction needs, uses spectral features to mine complex and fast hand movement features, and proposes a hand gesture recognition model. Extensive experiments in real environments show that DASIV achieves high-precision breathing monitor (Pearson correlation coefficient of 0.89) and hand gesture recognition (Precision of 91.7%).**

***Index Terms*—directional acoustic sensing, intelligent vehicle, breathing monitor, hand gesture recognition**

I. INTRODUCTION

Motivation: With increasing driving distances and time, there is a growing concern for drivers' physical and mental health and driving experience, which demonstrates that a safe and intelligent driving environment is becoming increasingly important [1], [2]. As acoustic sensors are increasingly used in car vehicles, smart audio systems, while providing applications such as stereo sound and intelligent interaction, may cause interference to drivers, thus endangering driving safety.

Directional sound fields [3], [4], prototypes such as directional speakers [5], [6], have been extensively studied in different scenarios. Consider a scenario where the driver can hear what they want while breathing is monitored to ensure their health (Fig. 1). Meanwhile, the driver can interact with various functions in the vehicle through the directional speaker for more flexible control. Such directional speakers can simultaneously broadcast directional sound and sensing, meeting the needs of interference-free listening, health monitoring, and human-vehicle interaction in the vehicle.

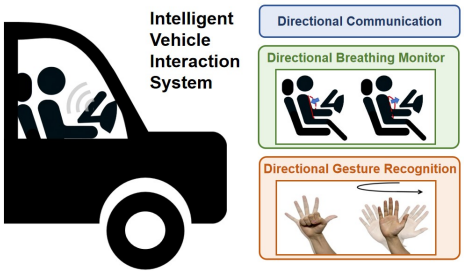


Fig. 1: Directional sensing and communication for the intelligent vehicle interaction system.

Unlike directional speakers, conventional acoustic sensing uses ordinary loudspeakers [7] that face several problems. Current acoustic-based breathing monitor systems [8], [9] face several common challenges associated with collecting breathing signals, including the need for a quiet environment and the subject to remain still. In a vehicle context, it is necessary to dynamically measure the driver's fine-grained breathing waveforms in a noisy, vibrating vehicle environment with passengers to monitor the driver's physical condition accurately. Similarly, some research suggests that hand gesture interaction is what consumers expect [10], which can effectively avoid accidents caused by distraction [11]. Current acoustic-based hand gesture recognition systems [12], [13] are challenging for vehicle environments due to complex interferences. Although there have been studies using beamforming for sensing [14], [15], the large size of beamforming via speaker array [16] makes it difficult to deploy in a vehicle.

Current directional speaker [5], [6] can provide a directional sound field, which is usable in vehicle environments because vehicle environments tend to have different listening needs for different people, and directional sound waves can be customized to the user's needs. However, these works did not leverage directional speakers to explore their potential capabilities that could be applied to sensing (e.g., breathing monitor and gesture recognition).

Directional speakers have a high directive nature, which can be exploited in sensing. Although conventional loudspeakers can achieve a directional sound field through beamforming technology [16], their large size makes them difficult to deploy

& Both authors contributed equally to the research.
* Corresponding author.

979-8-3503-6794-2/24/$31.00 ©2024 IEEE

in a vehicle. In comparison, an 8×8 directional speaker is only $9cm \times 9cm$ in size and is fully adapted to the vehicle environment. ***In the paper, we exploit this highly directional nature of directional speakers to achieve simultaneous high-precision directional user sensing in complex vehicle environments.***

Challenges: 1) Complexity of the vehicle environment: The vehicle environment has multi-path reflections from various objects and passenger interference, making it difficult to accurately extract the signals represented by breathing and hand gestures from the reflected signals. **2) Simultaneous sensing and communication**: The frequency band used for sensing must be inaudible to the user and not interfere with the original directional sound playback function of the directional loudspeaker, so the sensing signals need to be carefully designed to realize this simultaneous operation. **3) Driving Environment**: The vibration of a vehicle in motion can significantly affect the precision of the sensing. **4) Characterization of user hand gestures**: user hand gestures are often fast and subtle, which makes recognition challenging.

Solutions: 1) We used the directional nature of directional speakers to achieve directional sensing without losing the characteristics of the directional playback function and solve the interference problem in the complex vehicle environment. 2) We used high-frequency FMCW signals for sensing and circumvented the interference of differential frequency signals generated by air nonlinear effects to synchronize directional communication with sensing. 3) We used the signal phase to extract the thoracic motion caused by breathing, extracted the relevant frequency signals by bandpass and Gaussian filters, and fitted the real breathing signals with the bidirectional Long Short-Term Memory (Bi-LSTM) model. 4) We extracted the spatial dynamic features of user hand gestures using short-time Fourier transform (STFT) and achieved accurate hand gesture recognition by the attention-based ResNet. The main contributions of this work are summarized as follows:

- DASIV is the first high-precision directional sensing system based on directional speakers for synchronization with communications in a vehicle environment.
- The system captures weak chest movements based on phase features of FMCW signals and achieves a fine-grained breathing monitor based on Bi-LSTM.
- The system defines 10 interaction hand gestures under the driving state, extracts hand gesture features by STFT, and implements ResNet-based high-precision hand gesture recognition.
- We invited 10 drivers to validate DASIV in a vehicle environment, and the extensive experiments show that DASIV is capable of fine-grained breathing monitor (Pearson correlation coefficient of 0.89) and high-precision hand gesture recognition (precision of 91.7%).

II. Related Work

Breathing monitoring and hand gesture recognition can be broadly categorized into wearable device-based and contact-free methods. Wearable systems utilize physical characteristics like acceleration, moisture, and electrical potential changes to identify hand gestures and estimate breathing rate [17]–[19]. Recent optoelectronics-based systems [20] detect thoracic box expansion using voltage changes from reflected infrared light. Although these systems enhance user experience, they can interfere with daily routines.

Contact-free methods include vision-, RF-, and audio-based techniques. Vision-based approaches leverage cameras to segment gestures and estimate breathing rate using bag-of-features, color changes, and depth data [21]–[23]. However, they struggle in low-light conditions and pose privacy issues. RF-based methods initially targeted military applications and have now evolved into fine-grained detection tools, using signal processing and wireless location technologies [24], [25]. Despite their precision, they often require additional devices and are not suitable for vehicle environments. Audio-based approaches identify gestures and breathing rates by analyzing sound reflections from finger and chest movements. They can capture fine-grained motion with existing microphones [9]. The LLAP scheme [12] tracks finger movements to recognize characters drawn in the air. Privacy-sensitive sounds can be filtered out to retain only breathing signals [26], though differentiating non-gesture signals remains challenging in complex environments.

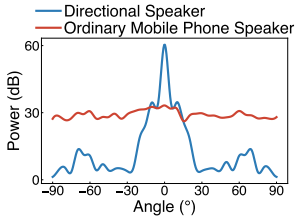


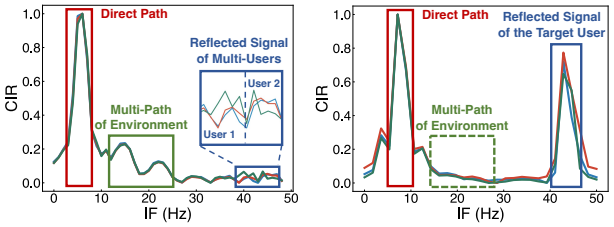
Fig. 2: Directional Sound Field. Directional speakers have a concentrated beam, while ordinary phone speakers have essentially the same intensity in all directions.

III. Preliminary

A. Directional Sound Field

The directional speakers prototype [5], [6] is based on the parametric array [27], a nonlinear acoustic mechanism that produces audible frequency difference waves via nonlinear ultrasonic interactions in air, known as air nonlinearity [28].

Thanks to the high directive of ultrasonic transducers, parametric arrays can function as directional speakers. Directional speakers can emit sound with concentrated intensity in a specific direction, producing a more focused sound field than ordinary loudspeakers, which usually emit sound omnidirectional. This difference in sound field distribution can be verified experimentally by analyzing the beam pattern received from the speakers. The beam pattern illustrates how the sound intensity varies with direction, providing insights into the directional characteristics of the speakers. Directional speakers exhibit sharp peaks in their beam patterns, indicating strong sound projection in specific directions, whereas ordinary mobile phone speakers typically display more uniform



(a) Mobile Phone Speaker (b) Directional Speaker

Fig. 3: The directional sound field can effectively suppress the multipath interference of the environment and the movement interference of other users and simultaneously improve the signal-to-noise ratio of the reflected signal from the target user.

distribution patterns, reflecting their omnidirectional nature. As shown in Fig. 2, we used mobile phone speakers and directional speakers to measure within a $180°$ range and found that directional speakers only spread between $70°$ and $100°$, while mobile phone speakers have almost the same effect at $180°$.

B. Feasibility of Directional Sensing

Traditional omnidirectional sensing is susceptible to interference of dynamic objects in the scene [29]. With increasing sensing range, subtle movements of the chest and hand are strenuous to be detected in the scene since reflection signals from the surrounding objects can be much stronger than these target signals [13]. In addition, multi-path reflection signals can distort received signals and contribute to phase measurement errors, especially in high precision applications [30]. Therefore, we introduce directional sensing based on directional speakers to solve these problems.

We verify the feasibility of directional sensing. Specifically, we compare the sensing between directional speakers and mobile phone speakers. As Fig. 3a shows, the peak value formed by the direct path is much higher than the peak value formed by the driver and passengers. At the same time, there will be many peaks formed by other stationary objects between the red part and the blue part. In addition, activities of the passenger and the driver will mix peaks of the two people, causing interference. In this way, the mobile phone speaker cannot locate the driver efficiently and clearly. In the Fig. 3b, the peak formed by the direct path and the peak formed by the driver's breathing is almost the same. At the same time, the peak formed by the interference of other items and passengers is completely eliminated.

IV. DASIV FRAMEWORK

A. Sensing via Directional Speaker

1) Signal Modulation: To achieve simultaneous directional communication and sensing, the modulation of the audible and FMCW signals needs to be addressed. Directional communication based on air nonlinearities utilizes squared (second order) terms, while higher order terms are negligible. A basic modulation scheme is amplitude modulation [31]. For a low-frequency audio signal $u(t)$ and a carrier frequency f_c, the modulated signal is: $s(t) = u(t)\cos(2\pi f_c t) + \cos(2\pi f_c t)$. Human hearing filters out high frequencies, resulting in the following low-frequency signal via the square term:

$$s_{low}(t) = \frac{a_2}{2}(2u(t) + u^2(t) + 1) \tag{1}$$

This allows ultrasound-based audio reproduction of $u(t)$.

Then, to construct the sensing ability of the directional speaker, we need to embed the FMCW into the transmitted signal, which must satisfy two conditions: 1) the FMCW signal occupies a band that does not coincide with the carrier signal and the up-converted audio signal; 2) neither the FMCW signal itself, nor its differential frequency signal with the carrier can be heard. Then, we explored the frequency bands that can be used by FMCW.

From the frequency response curve Fig. 5, it can be seen that the frequency band in which the communication signal is embedded is $36kHz - 44kHz$, considering the intensity requirements of the differential frequency audible signals. For FMCW signals, the optimal frequency range for an effective response is between $60kHz$ and $90kHz$. Within this range, the most favorable response occurs in the $70kHz - 80kHz$ band. Meanwhile, considering that the frequency resolution is affected by the bandwidth, a narrower bandwidth will result in a lower frequency resolution. Therefore, we decided to choose a bandwidth of $10kHz$ and set the frequency range up to $70kHz - 80kHz$.

2) Primer of FMCW: FMCW is characterized by its frequency that changes linearly over time. This signal is transmitted towards a target and the reflection is received back. By comparing the frequency differences between the transmitted and received signals, we can calculate the time delay of the signal propagation. Then we can estimate the distance.

A brief triangular wave FMCW is shown in Fig. 6. The transmitted signal of the FMCW can be expressed as

$$x_{tx}(t) = Acos(\phi(t)) = Acos(2\pi(f_0 t + \frac{kt^2}{2}) \tag{2}$$

Where f_0 is the starting frequency of the chirp signal, B is the bandwidth, A is the amplitude of the signal, T is the width of the chirp signal pulse, and $k = \frac{B}{T}$is the slope of the frequency change. The instantaneous phase of the transmitted signal is denoted as

$$\phi(t) = 2\pi \int_0^t f(t)dt = 2\pi(f_0 t + \frac{kt^2}{2}) \tag{3}$$

Assuming the distance between the target object and the sensing device is R, the time it takes for the signal emitted from the sensing device to reach the target object and reflect back to the sensing device is $t_R = 2R/c$, c is the sound velocity. So the received signal can be represented as $x_{rx}(t) = A_R cos(\phi(t - t_R))$, where A_R is the received signal amplitude. Next, multiply the received reflection signal with the standard FMCW transmission signal to obtain a mixed signal, which is the intermediate frequency signal $S_{if}(t) = x_{tx}(t) \cdot x_{rx}(t)$. And we can write the low frequency component of S_{if} is

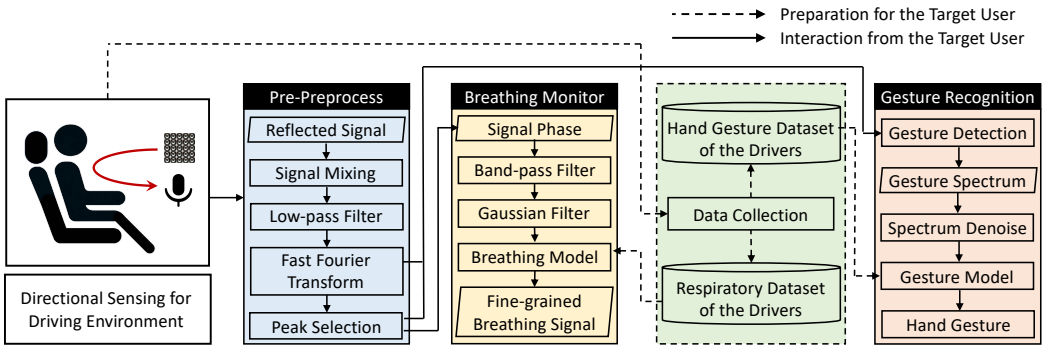


Fig. 4: Overview of DASIV, including signal transmission and collection, data processing, model training and prediction.

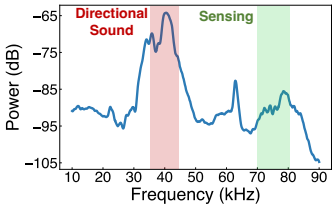


Fig. 5: Based on the frequency response curve of the ultrasonic oscillator in the directional speaker, the system sets the modulation frequency of the communication signal to $36kHz-44kHz$ and the modulation frequency of the FMCW signal to $70kHz - 80kHz$ to synchronize the directional sensing with the communication.

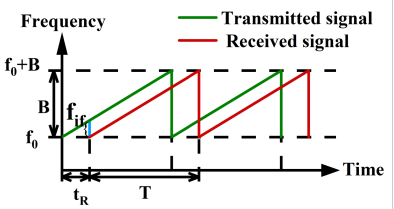


Fig. 6: FMCW transmitted and received signals

$f_{if} = \frac{2kR}{c} = \frac{2BR}{cT}$, if we know the f_{if} then we can get the distance R is

$$R = \frac{cf_{if}}{2k} \tag{4}$$

From the above equation, the distance resolution ΔR corresponds to the minimum frequency resolution Δf_{if}, which has the following relationship

$$\Delta R = \frac{c\Delta f_{if}}{2k} = \frac{c}{2B} \tag{5}$$

When we obtain the FFT spectrogram of f_{if}, a clear change curve can be seen on the spectrogram when the frequency difference Δf_{if} of a distance change is $\frac{1}{T}$. The bandwidth of DASIV FMCW is $10kHz$, and the sound speed $c \approx 343m/s$. Our distance resolution in the spectrogram is $1.72cm$.

B. Breathing Monitor

Breathing Monitor is one of the most important vital signs. The breathing wave can express the driver's physical condition. So it can provide a strong reference for physical health and driving safety.

Typically, the breathing process comprises inhalation and exhalation. During inhalation, as shown in Fig. 7(a), the diaphragm contracts and the intercostal muscles lift, expanding the chest. Conversely, during exhalation as depicted in Figure 7(b), the muscles relax and the chest returns to its resting position. The chest displacement caused by breathing typically ranges from approximately $1mm - 5mm$.

1) Phase-based Breathing Monitor: In the previous section, we provided a detailed introduction to the principle of direct speakers and FMCW signals. However, our task is not to do ranging. We want to monitor breathing, which needs us to discuss whether the resolution of FMCW sensing is suitable for current work.

As can be seen from the equation 5, if we use the spectrogram directly to estimate breathing waves. The distance variation of chest undulations caused by breathing will be submerged in the spectrum of f_{if}. It will appear only in a straight line in the spectrum. Therefore, we need more fine-grained measurement methods to obtain breathing waves.

Assuming that the movement of the human chest caused by breathing is Δd, f_{if}can be represented as $\frac{AA_R}{2}(cos(\frac{4\pi kRt}{c} + \frac{4\pi f_0 \Delta d}{c}))$, where $\frac{4\pi f_0 \Delta d}{c}$ is the phase change of the intermediate frequency signal, with a variation of $0 - 2\pi$.

As we know the displacement of the human chest during breathing is usually about $5mm$. Under normal circumstances, the breathing rate of an adult is between 12 - 20 breaths per minute [32]. Considering higher resolution, if the breathing rate is 10 breaths per minute. The amount of chest displacement that we need to distinguish for each chirp can be denoted as:

$$\delta d = \frac{\Delta d}{N \times \frac{60}{f_{breathmin} \times 2}} \tag{6}$$

Where every $5mm$ corresponds to one inhalation or one exhalation, chirp is $0.1s$ and N represents the number of chirps per second. And $f_{breathmin} = 10$, so the $\delta d = 0.16mm$.

Based on the chest displacement change of breathing, we can see obvious displacement. The phase change caused by every 0.16mm can be calculated as $\frac{4\pi f_0 \Delta d}{c} = 0.128\pi \approx 23.04°$, where the $f_0 = 7kHz$, $c \approx 343m/s$. This represents phase of f_{if}, DASIV can achieve the necessary resolution for effective breathing monitor.

2) Preprocessing of Breathing Monitor: In breathing monitor, we still face two challenges: 1) How to prevent anomalies caused by hand gestures or other driver activities; 2) How to address the slow phase drift in long-term breathing monitor.

(a) Inhale (b) Exhale

Fig. 7: Breathing can cause displacement of the chest.

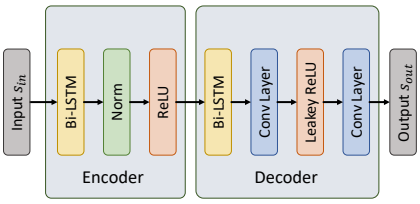


Fig. 8: Architecture of breathing model based Bi-LSTM.

For problem 1), we utilize a bandpass filter to extract signals at the frequency of breathing. Considering that the driver may sway back and forth within the vehicle, with a possible distance of $20cm$, corresponding to a frequency range of approximately $120Hz$ from the equation 4. Assuming the normal distance between the driver and the device is $80cm$, we extract the phase of the mid-frequency signal (ranging from $480Hz$ - $120Hz$ to $480Hz$ - $160Hz$) obtained through bandpass filter as the signal representing the breathing frequency. This approach helps to mitigate interference from hand gestures or other driver activities. For problem 2), we employ a Gaussian filter to address the slow phase drift.

3) LSTM-based breathing model: **Network architectures.** The architecture is shown in Fig. 8. The breath detection network utilizes a standard Encoder-Decoder structure to convert the input impulse sequence of length L_{in} into the corresponding standard breath sequence of length L_{out}. For the input sequence s_{in}, we first use a bidirectional LSTM [33] as an encoder to find and exploit long-range dependencies in the data and extract the features. The number of features in the hidden state is 64, and the number of layers is 2. Next, after the $BatchNorm - ReLU$ layers, we employ another bidirectional LSTM, identical in structure to the encoder, as a decoder to interpret the features. Additionally, we utilize a CNN to adjust their length to match the desired output length L_{out}. The convolutional layers both have L_{out} filters with a kernel size of 3×3. Next, we define the loss function for breath sequence prediction:

$$\mathcal{L} = \sum_{i=1}^{L_{out}} (s_{out}[i] - s_{gt}[i])^2 \tag{7}$$

where the s_{out} is the breath sequence output by the network and s_{gt} is the standard breath sequence obtained from the breath tape test. And we can employ the loss function to optimize and train our breath model.

C. Hand Gesture Recognition

1) Amplitude-based Hand Gesture Recognition: According to Sec. IV-A, we can obtain a spectrogram by performing FFT on f_{if}. The resolution of the spectrogram is $1.72cm$ in DASIV. Since distance changes caused by hand gestures exceed $5cm$, it is feasible for hand gesture recognition by spectrogram.

A critical problem is how to identify whether and how long a hand gesture occurs during continuous breathing monitor. We have introduced a detection algorithm to locate and segment audio segments where hand gestures occur. In the first step, we mixed the received signal with the transmitted signal and passed it with a low-pass filter. Then we performed FFT and conduct frequency analysis and select the two strongest peaks (One is the direct path, the other is the human body). Based on their frequencies, we designate the peak with the lower frequency as f_d and the peak with the higher frequency as f_p. During this process, if a third peak emerges between these two peaks and its peak surpasses the threshold S, we record a time window t_g until the amplitude of this peak falls below S. The audio segment corresponding to this time window t_g is identified as the segment where a hand gesture occurs.

2) Preprocessing of Hand Gesture: When the algorithm identifies the audio segments containing only hand gestures, we employ spectral subtraction to eliminate noise from the audio, thereby highlighting the features.

Consider the noise present in the leakage current, even if the user is not running another application. We denoise the signal using spectral subtraction [34] based on the noise samples taken at idle moments as follows:

$$\|Y(k)\|^2 = \|S(k)\|^2 - \alpha\|N(k)\|^2 \tag{8}$$

where k represents the frequency range of the band, $S(k)$ and $N(k)$ represent filtered signal and noise signal respectively. α is the ratio of the signal strength of each frequency corresponding to symbol 0 in the current channel to the signal strength of the noise sample.

The hand gesture audio segments are processed through spectral subtraction and transformed into spectrograms using STFT. These spectrograms are then uniformly cropped to serve as inputs for hand gesture recognition. Gesture design needs to be sufficient to not distract the driver. In DASIV, We have carefully designed 10 gestures. And we show the schematic diagram and feature map of each gesture in Fig. 9. We used 10 gestures, 8 of which are from [35]. We expanded 2 more gestures to improve the interactive function.

3) Gesture Model: **Network architectures.** The architecture is shown in Fig. 10. We use ResNet to accomplish this task, other networks such as VGGNet [36], LeNet-5 [37] have also been tried, but the effect is not as good as ResNet. And ResNet is accurate enough after adding attention module for our task. The hand gesture images are preprocessed into $3 \times 224 \times 224$ inputs and then processed through a convolutional layer, channel attention, spatial attention and max pooling layer, resulting in a dimension of $64 \times 64 \times 56$. The changed feature map is input into the residual module of ResNet18 [38]. After a series of residual operations such as convolution, regularization and identity mapping, a feature map of $512 \times 7 \times 7$ is obtained. It is further processed through channel attention, spatial attention and average pooling layer

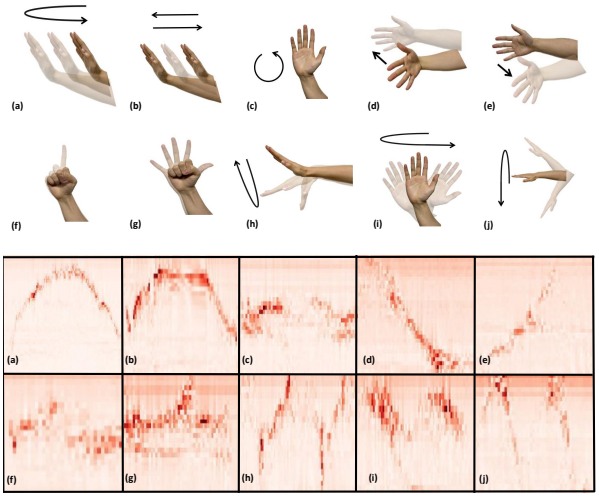


Fig. 9: Gesture schematic diagram and feature diagram, (a) Click. (b) Long press. (c) Circle draw. (d) Right sweep. (e) Left sweep. (f) One-finger fold. (g) Close and open. (h) Beckon. (i) Left and right wave. (j) Up and down sweep.

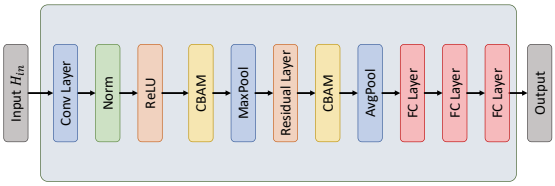


Fig. 10: Architecture of gesture model based ResNet.

to obtain a feature map of $512 \times 1 \times 1$. Finally, the feature map is used as the input for the fully connected layer. The classification results are output through four fully connected layers in sequence. The kernel size of the convolutional layers and pooling layers is 3×3, while $padding = 1$, $stride = 2$.

The loss function is Cross-Entropy Loss can be expressed as:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i log(p_i) \tag{9}$$

where N is the number of categories classified$y_i$is the actual target label, $p_i$is the predicted probability.

Attention architectures. Some of the hand gestures have relatively small direct differences and require more fine-grained classification. ResNet does not perform well in these types of hand gesture classification. Therefore, we introduced channel attention and spatial attention and connected these two types of attention in the order of channel attention and spatial attention to form a CBAM module [39]. So we added CBAM to ResNet18 to achieve more fine-grained gesture recognition.

The channel attention module in yellow part of Fig. 11 focuses on the channel of the given input image. The spatial attention module in green part of Fig. 11 focuses on feature position. These two modules are relatively independent, their use can be individually, in parallel, sequentially, or in reverse order into ResNet18.

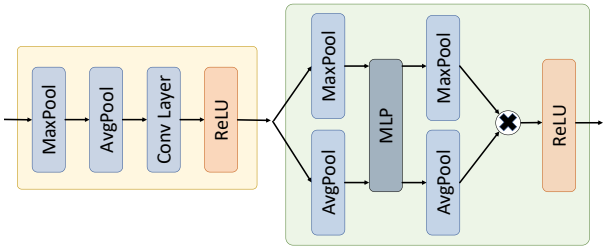


Fig. 11: Architecture of channel attention and spatial attention in CBAM.

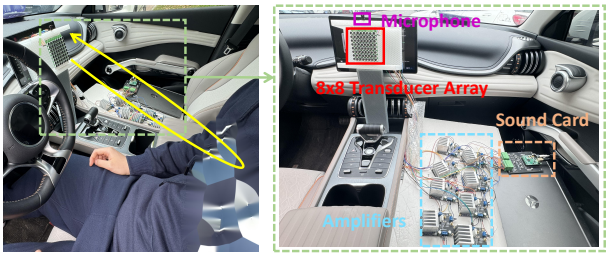


Fig. 12: DASIV deployment in the vehicle.

V. IMPLEMENTATION

A. Hardware

Our prototype includes an ultrasonic transducer array and a single microphone (in Fig. 12). The array features 8×8 EU16AOF40H12T transducers connected to an 8-channel audio source, such as a sound card linked to a laptop. Each channel, consisting of 8 ultrasonic transducers, is powered by a class D amplifier, the OPA541, capable of delivering up to $50W$ of power. The transducers have a central frequency of $40kHz$. As for the receiver, we employ a single microphone with a sampling rate of $192kHz$.

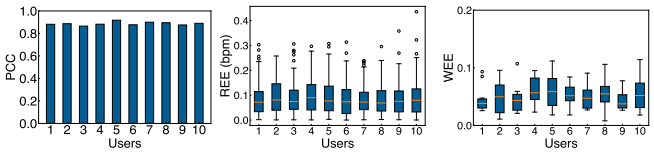
B. Dataset

We experimented with 10 drivers between the ages of 21 and 52, comprising 6 men and 4 women. Before data collection, we provided them with user manuals and privacy statements, which they were required to read carefully and sign consent forms. During the experiment, participants were instructed to complete the required operations in a safe driving environment. We collected a total of 20 hours of breathing curves with both DASIV data and breathing belts data (standard breathing wave). We required each experimenter to perform each gesture 60 times, and ensure that these 60 hand gestures were distributed at different times and in random order. Our final dataset has more than 24,000 sets (one set for 30 seconds) of breathing data and more than 6,000 sets of hand gesture data. Our dataset is divided into 70% training set, 20% validation set and 10% test set.

VI. PERFORMANCE EVALUATION

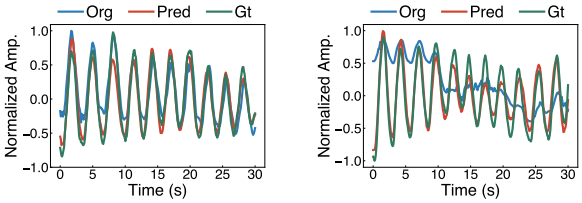
A. Evaluation Methodology

To assess the performance of DASIV, we consider 3 metrics to evaluate the experimental results of breathing monitor,



(a) PCC (b) REE (c) WEE

Fig. 13: Breathing Monitor Performance



(a) Driver sits still. (b) Driver is driving.

Fig. 14: Breathing Monitor Performance

and consider other 3 metrics to comprehensively evaluate the experimental results of hand gestures:

Rate Estimation Error (REE): The error of the estimated breathing rate (R_E) compared to the actual breathing rate (R_A), i.e., $|R_A - R_E|$.

Waveform Estimation Error (WEE): The error between predicted sequence s_{out} and actual sequence s_{gt}, i.e., $\sum_{i=1}^{L_{out}}(s_{out}[i] - s_{gt}[i])^2$.

Pearson Correlation Coefficient (PCC): r quantifies the strength and direction of a linear relationship between two continuous sequences s_{out} and s_{gt}, and can be calculated as:

$$r = \frac{\sum(s_{out}[i] - \bar{s_{out}})(s_{gt}[i] - \bar{s_{gt}})}{\sqrt{\sum(s_{out}[i] - \bar{s_{out}})^2 \sum(s_{gt}[i] - \bar{s_{gt}})^2}} \quad (10)$$

where $s_{out}[i]$ and $s_{gt}[i]$ are the individual elements of sequences s_{out} and s_{gt}, respectively. $\bar{s_{out}}$ and $\bar{s_{gt}}$ are the means of sequences s_{out} and s_{gt}, respectively.

Precision: the proportion of true positive predictions among all positive predictions.

Recall: the proportion of true positive predictions among all actual positive instances.

F1_score: the weighted average of precision and recall:

$$F1\_score = \frac{2 \times (precision \times recall)}{precision + recall} \quad (11)$$

B. System Performance

1) Breathing Monitor Performance: As shown in Fig. 13, we present an evaluation of the overall efficacy of the breath model for each driver across a range of breathing rates. In Fig. 13a, the Pearson correlation coefficients of the predicted and standardized breathing waveforms for various drivers are depicted. PCC of all the 10 experimenters(Sec. V-B) surpassing 0.85. This signifies a robust positive linear correlation between the predicted and standardized breathing waveforms. Fig. 13b illustrates the breathing rate estimation error, consistently below 0.4 bit per minute. Fig. 13c demonstrates the waveform estimation error, indicating that the disparities between the predicted and standard waveforms are less than 0.15. We can see in Fig. 14 that the driver is seated still, the three breathing waveforms are highly overlapped. When the driver is active, although the phase-based respiratory waveform become messy, the DASIV breathing waveform is still highly overlapped with the ground-truth.

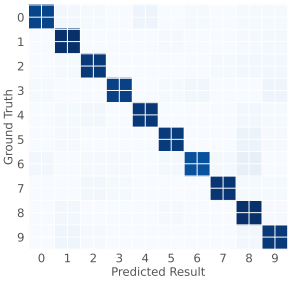
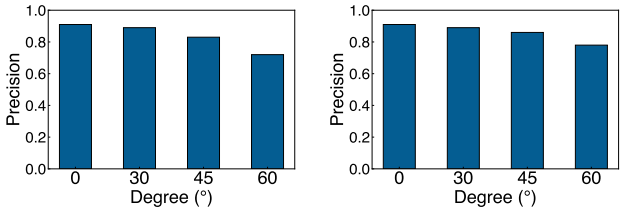


Fig. 15: Hand Gesture Recognition Performance.



(a) Bias of horizontal (b) Bias of Vertical

Fig. 16: Impact of user bias.

2) Hand Gesture Recognition Performance: Hand Gesture Recognition Performance of our DASIV system was evaluated by collecting 6,000 hand gestures from 10 drivers in the vehicle. The Fig. 15shows the confusion matrix on ten hand gestures. Numbers $1-10$ correspond to $(a)-(j)$ in Fig. 9 respectively. On the test set, the overall hand gesture recognition $F1\_score$ value is 92.3%, with an average $Precision$ 91.7%, indicating the feasibility and applicability of DASIV in achieving hand gesture interaction in the vehicle environment.

3) Impact of Bias: We assessed the impact of user hand gesture bias on the hand gesture recognition performance of directional speakers. We test the bias performance in both horizontal and vertical directions in Fig. 16. We observed that the performance under horizontal bias was lower than under vertical bias. Our designed hand gestures rely more on horizontal sensing abilities (e.g., left and right impact bias significantly). When the horizontal angle is less than $45°$, the precision exceeds 82.1%. The bias in the vehicle space is within the range of $45°$, which shows that DASIV can achieve high-precision hand gesture recognition under bias.

VII. DISCUSSION & CONCLUSION

DASIV can obtain high-precision respiratory waveforms. But DASIV needs to be able to accurately identify the driver's physical status, such as fatigue, difficulty breathing, etc., in the future. Besides, there are some unresolved issues encountered during hand gesture interaction. Firstly, certain habitual hand gestures of drivers may be mistakenly identified as our target hand gestures. Research on how to accurately distinguish target hand gestures is significant in the future investigation.

Secondly, users exhibit varying habits when performing hand gestures, such as the orientation of their palms, which can result in differences in the size of the reflective surfaces, which will lead to confusion in some hand gestures with similar features. This also needs to be solved in the future. We use ResNet as the basic neural network for hand gesture recognition. Network innovation is not our focus, but we can try to achieve a lighter with higher precision network in the future. In addition, DASIV is directional and supports an angle range of $\pm 60^{\circ}$, which is supportable in most scenarios. However, during the experiment, special driving habits such as leaning forward or back and adjusting the seat too high or too low can affect system performance. We recommend that users adjust the directional speakers to match their driving style.

In this paper, we introduce DASIV, which utilizes the highly directional nature of ultrasonic signals to achieve fine-grained directional acoustic sensing in the vehicle. DASIV enables synchronized directional acoustic communication. Extensive experimental verification showed that DASIV can achieve high-precision and robust respiratory monitoring and gesture recognition.

ACKNOWLEDGMENT

This work is supported in part by the NSFC (61936015, 62072306) and Shanghai Key Laboratory of Trusted Data Circulation, Governance and Web3.

REFERENCES

[1] R. Steinbach and B. C. Tefft. American driving survey: 2022, 2023.
[2] Lizzie Nealon. 2023 car ownership statistics, 2023.
[3] DW Robinson and LS Whittle. The loudness of directional sound fields. *Acta Acustica united with Acustica*, 10(2):74–80, 1960.
[4] Haoran Xue, Yihao Yang, and Baile Zhang. Topological acoustics. *Nature Reviews Materials*, 7(12):974–990, 2022.
[5] Holosonics. Holosonics audio spotlight series. https://www.holosonics.com/. Accessed on May 1, 2023.
[6] Focusonics. Focusonics directional speakers. https://www.focusonics.com/. Accessed on May 1, 2023.
[7] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. Acoustic-based sensing and applications: A survey. *Computer Networks*, 181:107447, 2020.
[8] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. Contactless infant monitoring using white noise. In *ACM MobiCom*, 2019.
[9] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *ACM MobiSys*, 2019.
[10] Francisco Parada-Loira, Elisardo González-Agulla, and José L Alba-Castro. Hand gestures to control infotainment equipment in cars. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 1–6. IEEE, 2014.
[11] Motorcycle Helmet Use Laws. Traffic safety facts.
[12] Wei Wang, Alex X Liu, and Ke Sun. Device-free gesture tracking using acoustic signals. In *ACM MobiCom*, 2016.
[13] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. Room-scale hand gesture recognition using smart speakers. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 462–475, 2022.
[14] Gaetano Licitra, Francesco Artuso, Marco Bernardini, Antonino Moro, Francesco Fidecaro, and Luca Fredianelli. Acoustic beamforming algorithms and their applications in environmental noise. *Current Pollution Reports*, 9(3):486–509, 2023.
[15] Haocheng Hua, Jie Xu, and Tony Xiao Han. Optimal transmit beamforming for integrated sensing and communication. *IEEE Transactions on Vehicular Technology*, 2023.
[16] Ang Shiming and Chen Yaosen. Sound control using speaker array. *A university wide programme, URECA or Undergraduate Research Experience on CAmpus*, page 69.
[17] Jiahui Wu, Gang Pan, Daqing Zhang, Guande Qi, and Shijian Li. Gesture recognition with a 3-d accelerometer. In *Ubiquitous Intelligence and Computing: 6th International Conference, UIC 2009, Brisbane, Australia, July 7-9, 2009. Proceedings 6*, pages 25–38. Springer, 2009.
[18] Amy Sarah Ginsburg, Jennifer L Lenahan, Rasa Izadnegahdar, and J Mark Ansermino. A systematic review of tools to measure respiratory rate in order to identify childhood pneumonia. *American journal of respiratory and critical care medicine*, 197(9):1116–1127, 2018.
[19] John Hampton and Joanna Hampton. *The ECG Made Easy E-Book: The ECG Made Easy E-Book*. Elsevier Health Sciences, 2019.
[20] Daniel Llamas-Maldonado, Grace Leslie, Juan Alfonso Salazar-Torres, Adriana Del Carmen Téllez-Anguiano, and Miguelangel Fraga-Aguilar. Design of a physiological parameter monitoring system, implementing internet of things communication protocols by using embedded systems. In *IEEE ROPEC*, 2022.
[21] Nasser H Dardas and Nicolas D Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE TIM*, 2011.
[22] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM TOG*, 31(4):1–8, 2012.
[23] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Multi-sensor system for driver's hand-gesture recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.
[24] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. Smart homes that monitor breathing and heart rate. In *ACM CHI*, pages 837–846, 2015.
[25] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. Whole-home gesture recognition using wireless signals. In *ACM MobiCom*, 2013.
[26] Kaiyuan Hou, Stephen Xia, and Xiaofan Jiang. Buma: Non-intrusive breathing detection using microphone array. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pages 1–6, 2022.
[27] F Joseph Pompei. *Sound from ultrasound: The parametric array as an audible sound source*. PhD thesis, Massachusetts Institute of Technology, 2002.
[28] David G Crighton. Model equations of nonlinear acoustics. *Annual Review of Fluid Mechanics*, 11(1):11–33, 1979.
[29] Fadel Adib, Zachary Kabelac, and Dina Katabi. {Multi-Person} localization via {RF} body reflections. In *USENIX NSDI*, 2015.
[30] Michael S Braasch. Multipath. *Springer handbook of global navigation satellite systems*, pages 443–468, 2017.
[31] Masahide Yoneyama, Jun-ichiroh Fujimoto, Yu Kawamo, and Shoichi Sasabe. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *JASA*, 73(5):1532–1536, 1983.
[32] WC Adams. Measurement of breathing rate and volume in routinely performed daily activities. *Final Report Contract*, (A033-205), 1993.
[33] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
[34] S. Kamath and P. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE ICASSP*, volume 4, 2002.
[35] Landu Jiang, Mingyuan Xia, Xue Liu, and Fan Bai. Givs: Fine-grained gesture control for mobile devices in driving environments. *IEEE Access*, 8:49229–49243, 2020.
[36] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019.
[37] Ahmed El-Sawy, Hazem El-Bakry, and Mohamed Loey. Cnn for handwritten arabic digits recognition based on lenet-5. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2*, pages 566–575. Springer, 2017.
[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.
[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.