AeroSense: Sensing Aerosol Emissions from Indoor Human Activities

BHAWANA CHHAGLANI, University of Massachusetts Amherst, USA
CAMELLIA ZAKARIA, University of Toronto, Canada
RICHARD PELTIER, University of Massachusetts Amherst, USA
JEREMY GUMMESON, University of Massachusetts Amherst, USA
PRASHANT SHENOY, University of Massachusetts Amherst, USA

The types of human activities occupants are engaged in within indoor spaces significantly contribute to the spread of airborne diseases through emitting aerosol particles. Today, ubiquitous computing technologies can inform users of common atmosphere pollutants for indoor air quality. However, they remain uninformed of the rate of aerosol generated directly from human respiratory activities, a fundamental parameter impacting the risk of airborne transmission. In this paper, we present *AeroSense*, a novel privacy-preserving approach using audio sensing to accurately predict the rate of aerosol generated from detecting the kinds of human respiratory activities and determining the loudness of these activities. Our system adopts a privacy-first as a key design choice; thus, it only extracts audio features that cannot be reconstructed into human audible signals using two omnidirectional microphone arrays. We employ a combination of binary classifiers using the Random Forest algorithm to detect simultaneous occurrences of activities with an average recall of 85%. It determines the level of all detected activities by estimating the distance between the microphone and the activity source. This level estimation technique yields an average of 7.74% error. Additionally, we developed a lightweight mask detection classifier to detect mask-wearing, which yields a recall score of 75%. These intermediary outputs are critical predictors needed for *AeroSense* to estimate the amounts of aerosol generated from an active human source. Our model to predict aerosol is a Random Forest regression model, which yields 2.34 MSE and 0.73 r^2 value. We demonstrate the accuracy of *AeroSense* by validating our results in a cleanroom setup and using advanced microbiological technology. We present results on the efficacy of *AeroSense* in natural settings through controlled and in-the-wild experiments. The ability to estimate aerosol emissions from detected human activities is part of a more extensive indoor air system integration, which can capture the rate of aerosol dissipation and inform users of airborne transmission risks in real time.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → Machine learning; • **Hardware** → *Sensor applications and deployments.*

Additional Key Words and Phrases: Aerosol Sensing, Audio Sensing, Privacy, Mobile Health

ACM Reference Format:
Bhawana Chhaglani, Camellia Zakaria, Richard Peltier, Jeremy Gummeson, and Prashant Shenoy. 2024. *AeroSense*: Sensing Aerosol Emissions from Indoor Human Activities . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 49 (June 2024), 30 pages. https://doi.org/10.1145/3659593

1 INTRODUCTION

Since humans spend over 90% of their time indoors, indoor air quality significantly impacts their health and wellness. Many studies have shown that poor indoor air quality triggered by common indoor air pollutants,

Authors' Contact Information: Bhawana Chhaglani, University of Massachusetts Amherst, USA, bchhaglani@cs.umass.edu; Camellia Zakaria, University of Toronto, Canada, camellia.zakaria@utoronto.ca; Richard Peltier, University of Massachusetts Amherst, USA, rpeltier@umass.edu; Jeremy Gummeson, University of Massachusetts Amherst, USA, gummeson@cs.umass.edu; Prashant Shenoy, University of Massachusetts Amherst, USA, shenoy@cs.umass.edu.

ACM 2474-9567/2024/6-ART49
https://doi.org/10.1145/3659593

including dust, mold, fine particulate matter of less than 2.5 microns (PM2.5), or volatile organic compounds (VOC), worsens allergy symptoms and asthma. More recently, there has been a heightened concern about poor indoor air quality through airborne transmission diseases, where viral loads from Flu (influenza) and COVID-19 spread from human-expelled aerosol droplets that remain in the indoor air [1–5]. These concerns have highlighted the need for sensing aerosol emissions and transmission risk in indoor environments.

Aerosol emissions in the air can be directly sensed by using advanced particle sensing technologies such as Condensation Particle Counters [6]. However, such particle counters are highly specialized, very expensive instruments that are not designed for deployment and everyday use at scale in indoor settings. An alternative to the direct sensing methods is indirect sensing of aerosol emissions, beginning with discerning everyday human activities that are known to contribute to aerosol generation. For example, human activities such as coughing, sneezing, and, very simply, speaking close to others can produce high numbers of aerosol droplets and contribute to the spread of airborne diseases [7]. While sensing of such human activities can be done using modalities such as audio [8–10], translating these activities to aerosol emissions remains unstudied. Recently, Chhaglani et al. [11] proposed the concept of using audio sensing to estimate aerosol emissions, but this work focused on the feasibility of the idea and articulating research challenges that needed to be overcome for a practical system, while leaving the design and evaluation of the system to future work.

In this paper, we present *AeroSense*, which uses a single sensing modality, namely audio, to offer a novel low-cost and privacy-preserving sensing approach to estimate the rate of aerosols generated in an indoor space from detecting human activity. Since audio sensing has the inherent disadvantage of capturing human speech, our system employs a *privacy-first* design methodology to ethically deploy such systems for large-scale monitoring without violating users' privacy, especially in community spaces. That is, rather than using raw audio signals, the system uses a small set of audio features that never permit reconstructing the original audio.

At its heart, *AeroSense* detects several aerosol-generating human activities such as coughing, sneezing, and talking. However, the problem of translating an *activity type* into aerosol generated from that activity comes with non-trivial challenges. Specifically, the number and size of aerosols expelled are influenced by the *loudness* of voice, which has been reported in prior studies as a significant factor [10], in addition to the *number* of human sources expelling aerosol in the same room. Another critical factor significantly reducing aerosol transmission between humans is *mask wearing* [12]. While there has been recent work on how mask-wearing impacts the acoustic characteristics of a person's voice [13] and detection of mask-wearing using the full audio spectrum [14], such approaches do not perform aerosol estimation and nor do they preserve the privacy of human speech. In designing, implementing, and evaluating *AeroSense*, our paper makes the following contributions:

First, we present a low-cost sensing mechanism that adopts a privacy-first design to estimate aerosol emissions. Our solution is deployed as a single audio modality that strategically captures a small set of non-reconstructible audio features to determine the factors that make up the rate of aerosol expelled from humans. Second, we develop machine learning and statistical models to detect activity types, discern mask wearing and distinguish between electronic voices (e.g., zoom meeting participants) from physically present speakers. Further, we present estimation techniques for activity loudness and aerosol sources, which together serve as critical parameters for predicting aerosol emissions in an indoor space. We implement a full prototype of our *AeroSense* using a Raspberry Pi 3B+ and two omnidirectional microphone arrays to predict all of the above factors and localize the activity by measuring direction-of-arrival, thus allowing us to calculate the distance and direction of the human source. Our prototype uses a privacy-preserving feature extraction pipeline at the device layer, in which features serve as input to higher-level prediction and statistical models. Fourth, we demonstrate the robustness of *AeroSense* by providing a detailed evaluation of *AeroSense* in a cleanroom, controlled real-world, and in-the-wild settings. Our results from testing a Random Forest regression model yield 2.34 MSE and 0.73 r^2 value. The cleanroom allows us to precisely predict aerosol droplets without accounting for other particles expelled in the environment. Our model accurately predicts the activity types in controlled settings at 85% recall on average.

Additionally, our activity level estimation technique had 7.74% error on average. Our mask detection model had 75% recall. Our electronic voice detection technique can effectively detect voice liveness with up to 80% accuracy. Our in-the-wild experiments demonstrate how *AeroSense* performs under indoor conditions with lower and higher risk of airborne transmission. We provide the source code of *AeroSense* as well as our audio and aerosol datasets to the research community.

2 BACKGROUND

This section provides background on aerosol sensing.

2.1 Indoor Air Quality

The topic of indoor air quality has, by and large, emphasized the presence and removal of indoor particulates (e.g., PM2.5) and volatile organic compounds (VOCs); recently there has been renewed interest in reducing airborne transmissions and virus spread between and among occupants [1]. Airborne transmitted diseases are more likely to occur in poorly ventilated and crowded settings. The droplets expelled due to talking, coughing, or sneezing can persist in the air from seconds to hours, depending on particle size, room volume, and air currents. Tiny droplets will evaporate into the indoor air and can last for more extended periods [15]. Nonetheless, the half-life of aerosol droplets' has been found to drastically decrease to 6 minutes with increased ventilation [16]. The exposure between persons to these droplets creates potential transmission routes. Since the coronavirus pandemic, mask-wearing and social distancing are equally stressed to dampen the expulsion and transmission of aerosol droplets [2–4]. Surgical masks and KN95 respirators have reportedly reduced outward aerosol particle emissions rates by 90% and 74% on average when speaking and coughing, compared to unmasking [12]. Our efforts in improving indoor air quality focus on measuring aerosol emissions, specifically from human expulsion in everyday respiratory-typed activities.

2.2 Ventilation Sensing

A building's ventilation system is designed to remove stale air from indoor spaces and replace with fresh air trhough circulation, thereby eliminating indoor pollutants (e.g., dust, allergens, and VOC particles) as well as airborne viral loads. Sensors integrated in building management systems can monitor airflow using air flow meters, pressure sensors, and vane anemometers through ducts and vents. However, these systems are typically hard-wired and require commissioning to install and need to be calibrated by facility managers. A recent work by Chhaglani *et al.* found airflow sensing was possible using only the existing sensors in smartphones to provide accessibility for everyday users [17]. Today, HVAC systems are coupled with real-time occupancy and CO_2 monitoring to ensure optimum air circulation based on safe CO_2 thresholds [18, 19]. While CO_2 generated from human occupancy is highly relevant as a critical factor in modulating indoor ventilation, these measures cannot infer aerosol amounts expelled into the air from occupants' activities. In such cases, the estimation of aerosol emissions offers more precision because the types and levels of activities are more related to aerosol generation than CO_2 generation rates [20]. Further, medical experts have found conflicting results in the correlation between elevated CO_2 levels and SARS-CoV-2 transmission risk [21] versus airborne propagation of typical respiratory problems [22].

2.3 Aerosol Emissions from Human-Respiratory Activities

By definition, aerosol particles are "very finely subdivided liquid, or solid particles dispersed in and surrounded by a gas" [23]. These particles come from varied sources, including naturally occurring from humans. For example, on the macro level, commuting is an emission source of aerosol particles associated with urban smog [24]. At the micro-scale, these particles are attributed to a person's virus-containing body secretions (e.g., respiratory

droplets) and can be transmitted through everyday activities [25]. Respiratory aerosol droplets released from humans while speaking, singing, exercising, coughing, and sneezing can be of various sizes. The largest droplets settle quickly due to gravity, while the smallest droplets can remain suspended in the air for minutes to hours and inhaled by a healthy person [26].

Table 1 lists the rate of aerosol particles generated from common human respiratory-typed activities [25]. A study specific to the recent coronavirus pandemic found that the leading cause for most super-spreader events is talking loudly [27]. By simply talking, Asadi *et al.* reported that, on average, the aerosol particles expelled from a person are larger and carry higher amounts of pathogens than breathing. Since particle emission rate in a speech is linearly correlated with amplitude (i.e., loudness), the impact of voice volume is an essential indicator of aerosol generation [10]. Inherently, a room full of talking adults will produce more aerosol [28, 29] than a room half-filled. One way to reduce aerosol generation into the air is mask-wearing [12]. Conversely, factors that help dilute aerosol concentrations in indoor air are the ventilation rate, air filtration standards, and room dimension [30, 31].

Table 1. Human activity and average aerosol emission reported in prior studies.

Activity	**Avg. Emission**	**Activity**	**Avg. Emission**
Breathing [10, 32]	<2 P/s	Quiet Talking (≈ 70 dB) [10]	6 P/s
Normal Talking (≈ 85 dB) [10]	14 P/s	Loud Talking (≈ 98 dB) [10]	53 P/s
Sneezing [33]	40000/Sneeze	Coughing [33]	3000/Cough

2.4 Aerosol Sensing and Audio

Since direct sensing of aerosols using particle counters is prohibitively expensive and such instruments are not designed for continuous ambient monitoring, indirect sensing aerosol emissions by monitoring human activities that produce emissions is a promising approach. Previous research has used audio sensing to detect human respiratory activities such as talking, breathing, and coughing [8]. However, such methods are typically based on capturing raw audio, containing human speech and leaks privacy. Further, the translation of these sensed activities to the amount of aerosols they generate remains an unaddressed gap. Finally, audio-based indirect aerosol sensing will not capture all aerosol emissions and will only capture those generated by humans through speaking, coughing, just to name a few. Given our goal of using aerosol emissions to monitor the risk of transmission of airborne diseases, this approach is well suited for our needs.

3 AUDIO SENSING FOR AEROSOL: MOTIVATION

Next, we motivate the AeroSense design and describe the key challenges.

3.1 Design Rationale

A key hypothesis of our work is that the level of aerosol generated by humans depends on their activities, which are correlated with the type and level of audio generated. To validate our hypothesis, we conducted a controlled experiment in a clean room, Using a condensation particle counter (CPC) [6]. Our setup, in Section 6, instructs a user to perform different respiratory activities to measure the audio produced by these activities using microphones and the rate of aerosol generated using CPC.

3.1.1 ***Relation between Activity Type, Aerosol and Audio.*** Figure 1 (top-left) shows the amount of aerosol the CPC measures during breathing, talking, coughing, and sneezing generated by a single user in a controlled setting. These activities generate increasing amounts of aerosol, which aligns with our intuition. Breathing

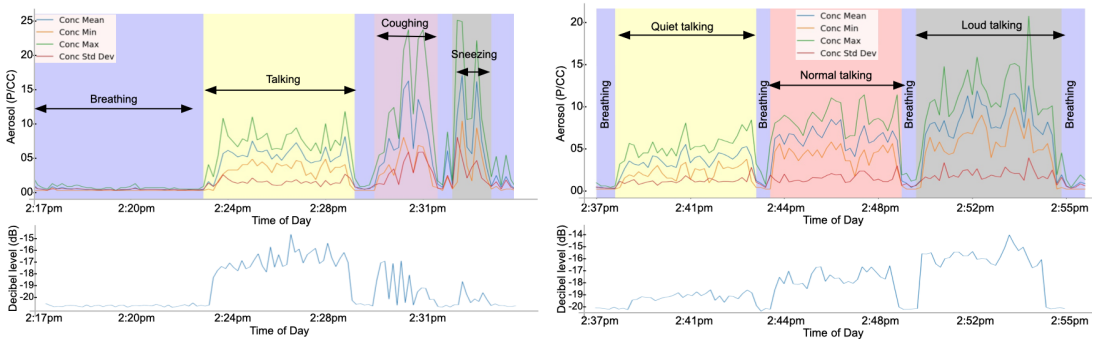


Fig. 1. (Left) Top shows aerosol varies with different activity types. Bottom shows audio signals varies with different activities. (Right) Top shows aerosol varies with different activity levels. Bottom shows audio signals varies with different activity levels.

generates less than 0.5 average aerosol particles per cubic centimeter (p/cc), while talking results in 2-13 p/cc. Coughing and sneezing can generate more than 20 p/cc on average. Figure 1 (bottom-left) shows the amplitude of the audio signal (in decibels) generated by these activities. Specifically, the audio pattern varies by activity, and the audio signals correlate with the aerosol emissions rate, with more concentrated aerosol-generating activities at higher audio levels. The findings from our controlled experiments are also supported by prior work [28, 29, 34–37], which show variations in aerosol generation during similar activities of talking, coughing, and sneezing. For this experiment, the flow rate of CPS is set to 1.5 L/min, which is $25cm^3/s$. We use this flow rate to find the number of particles per second, P/s = (p/cc) $*$ (cc/s), meaning the particle concentration of 5 p/cc is 5*25 P/s = 125 P/s. Using this calculation, we compute that breathing generates up to 12 P/s, speaking generates 50-325 P/s, and coughing and sneezing can generate more than 500 P/s. This observation is consistent with Table 1, where the experiments were performed with a flow rate of 1 L/min.

3.1.2 ***Relation between Activity Level, Aerosol and Audio.*** Figure 1 (top-right) shows the aerosol generation measured by the CPC when the user is talking at different loudness, ranging from soft to normal talking and shouting. As shown, the louder the speech, the greater the amount of aerosol generated. Soft speech generates less than 10 p/cc, while loud speech generates twice as much aerosol. Figure 1 (bottom-right) shows the corresponding audio in terms of its amplitude. As shown, the decibels of captured audio increases with speech loudness. This experiment confirms that even for a single aerosol-generating activity such as talking, the rate of aerosol generated is correlated with activity loudness and, thus, the audio level. Our results align with the findings of Asadi *et al.*, who reported that the particle emission rate of speech is linearly correlated with the amplitude of the vocalization [10].

Key Takeaway Together, these preliminary findings confirm that audio signals from human respiratory activities and their respective loudness are correlated with the rate of aerosol expelled. These insights suggest that audio sensing is viable for determining aerosol generation in indoor spaces.

3.2 Key System Challenges

We describe several challenges to address for a practical aerosol estimation solution using audio sensing.

3.2.1 ***Privacy-preserving Audio Sensing.*** Figure 2 shows the audio frequency spectrogram of different respiratory activities: coughing, sneezing, speaking, and speaking with mask. It can be seen that the spectrogram shows distinct features for each of these activities. A key design goal for *AeroSense* is to sense the parameters

needed to estimate aerosol expelled by humans in a privacy-preserving manner, meaning raw audio cannot be directly used in our system, and we must instead use non-reconstructible features of the audio to perform all sensing and detection tasks. We argue that such features can be derived from the sensed audio either in hardware or at the operating system level (e.g., in the audio drivers) so that applications such as *AeroSense* never deal with raw audio data and only work with the extracted features. We note that some audio speech recognition systems use audio features such as Fast Fourier Transform (FFT) or Mel-frequency Cepstral Coefficients (MFCC) [8], but even these techniques can be used to reconstruct the original audio. Hence, *AeroSense* must work with a more limited set of audio features for privacy reasons.

3.2.2 ***Isolating Speech Features from Speech.*** Prior work has reported the effects of voicing and articulation on aerosol generation [9]. However, detecting speech content characteristics is challenging without full audio. As a workaround, *AeroSense* seeks to determine the activity loudness, which can measured by the amplitude of sound (decibel, dB). This feature, however, is insufficient by itself because it discounts the actual decibel level originating from the (human) source and only reports the received level. As shown in Figure 3, the signal peaks become increasingly dissimilar with distance. Overcoming this challenge necessitates our system to determine the distance between the source and the audio sensor. Our solution employs two omni-directional microphones to calculate the direction of arrival estimate and triangulate the source.

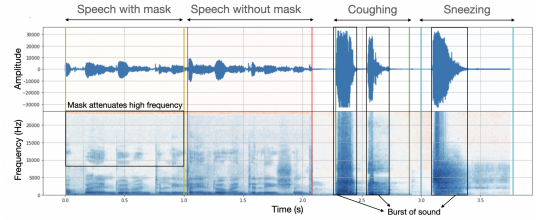


Fig. 2. Audio signal and its spectrogram for different activities.

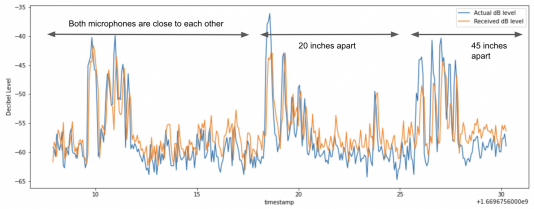


Fig. 3. Received decibel level decreases with distance between human source and audio sensor.

3.2.3 ***Mask Detection with Audio.*** Mask-wearing has become standard practice for preventing the spread of airborne viruses (e.g., influenza). There has been some advancement in mask detection using images [38] and audio [14]. Notably, the prior audio-based method requires a full spectrogram and cannot be applied to our system. Overcoming this challenge necessitates our system to design a lightweight mask detection model using a limited set of non-reconstructible features as its solution.

3.2.4 ***Electronic Voice Detection.*** Since office environments will frequently include electronic voices (e.g., speakers on Zoom conference calls, during playback of recorded lectures), incorrectly classifying these voices as physically present speakers will result in overestimating aerosol emissions. Consequently, our system must differentiate between actual and electronic speakers whenever it detects human speech in the environment.

3.2.5 ***Simultaneous Activity Occurrences.*** One or more persons performing different activities can occupy an indoor space. To effectively detect multiple activities in an indoor environment, our system must employ concurrent ML models to detect and distinguish between everyday human activities they are performing. Our system must also determine the number of human sources present, which is done using audio's direction of arrival (DOA). Since humans can move around a confined space when performing activities, we must also avoid double counting users when using the direction of arrival methods, which we achieve using a sliding window.

4 *AEROSENSE* SYSTEM DESIGN

In this section, we present an overview of our system and the design details of individual components.

4.1 System Overview

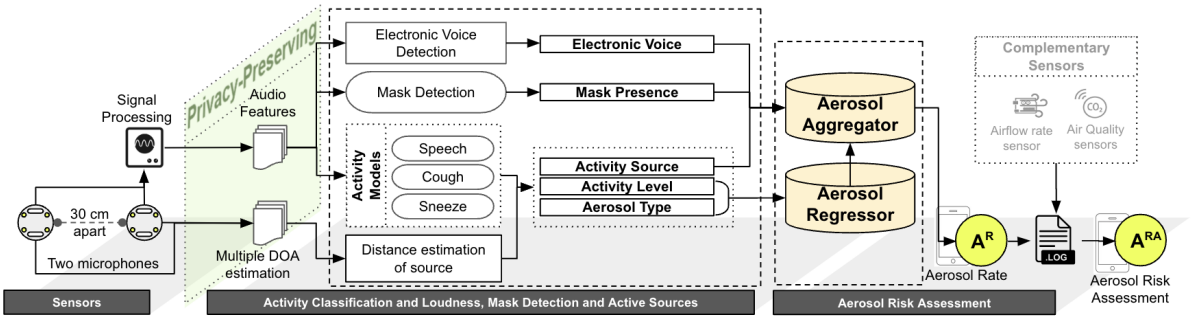


Fig. 4. The *AeroSense* system predicts the rate of aerosol generated from human activities, A^R, using non-reconstructible audio features. With supplementary parameters and complementary sensors to capture the rate of aerosol dissipation, the system supports providing indoor risk assessment, A^{RA}.

Figure 4 shows the design of our proposed approach, *AeroSense*. Our approach to estimating aerosol emissions is based on the use of a single sensing modality — audio. At the most basic level, *AeroSense* must determine two key parameters: *activity_type* and *activity_loudness*. Additionally, we consider detecting *mask_presence*, *active_sources*, and *electronic_voice* as the additional parameters for accurate aerosol estimation.

As illustrated in Figure 4, *AeroSense* utilizes two omnidirectional microphone arrays. Since audio data is highly invasive by nature and often raises privacy concerns, our system adopts a privacy-by-design approach, requiring the system to utilize different signal processing techniques within the device driver to extract numerous features and discard the full-spectrum raw audio. Using only non-reconstructible features, the system employs a combination of machine learning (ML) models to systematically determine binary outcomes of the above parameters to estimate aerosol emissions. The accumulation of estimated aerosol, combined with secondary parameters of the rate of aerosol dissipation (using complementary indoor air quality sensors), can support the system in assessing airborne transmission risk, A^{RA}, over a more extended period.

Given a privacy-by-design approach, we extract the non-reconstructible features directly in hardware or at the operating system level (e.g., in the audio drivers) to prevent *AeroSense* from processing or storing raw audio data. Some audio speech recognition systems use audio features such as Fast Fourier Transform (FFT) or Mel-frequency Cepstral Coefficients (MFCC) [8]. However, these techniques can be used to reconstruct the original audio which means that *AeroSense* must work with a more limited set of audio features to ensure privacy.

4.2 Privacy-Preserving Features

Table 2 provides the full list of extracted privacy-preserving features. Audio signals from two microphones are acquired at a sampling rate of 16 kHz with each sample represented as 16-bit integer values. The sampled audio stream is then segmented into non-overlapping frames of 500 *ms*. Our work does not use raw audio data and only makes use of non-reconstructible features extracted at the device layer.

The base frequency of a speech signal (F_0), is defined as the frequency at which the vocal folds vibrate when voiced speech sounds are made, ranges between 80 to 255 Hz [39] (note: the voiced speech of a typical adult male will have a fundamental frequency from 85 to 155 Hz, and that of a typical adult female from 165 to 255 Hz). P_{ratio} is the ratio of energy at base frequency plus energy at first harmonic to the total energy of the entire segment[40]. We used L/H_{ratio} as a key feature for mask-wearing, as it would significantly decrease mean spectral

levels at high-frequency regions as shown in Figure 2 and the harmonics-to-noise ratio, HNR, which increases in the presence of a mask [41]. The burst of sound from coughing generates significant energy well into the 15 kHz range, as shown in Figure 2. As per prior work [42], cough is modeled utilizing features such as the mean decibel energy, (P_{avg}), mean decibel energy above 8 kHz ($P_{>8}$), and below 8 kHz ($P_{<8}$) from using a fast Fourier transform (FFT) algorithm. To characterize vowel segments, we followed prior work [9] in using formant features ($Formant_{1-5}$), which have proven to contain evidence of periodicities of vowels and, thus, distinguish one vowel sound from another [43]. Filter banks are an array of triangular filters that split the spectrum into different components. If the FFT size is 512, number of filter banks is 12, we get the center frequencies [308.92 , 617.85...3707.09], given a 16000 Hz sampling frequency.

Table 2. List of features for activity type classification.

Feature	**Description**	**Feature**	**Description**
F_0: Base Frequency	Average number of oscillations per second	HNR: Harmonic to noise ratio	Ratio between voice and noise-like components of a speech sound
L/H_{ratio}: Low to High ratio	Ratio of power below 1 kHz to power above 1 kHz	P_{ratio}: Power ratio	Ratio of energy sum of the base and the first harmonic to the entire segment
$P_{>8}/P_{<8}$: Power above or below 8 kHz	Mean decibel energy of the FFT coefficients above or below 8 kHz	P_{avg}: Mean Power	Mean decibel energy of the entire FFT
RMS: Root mean square energy	Time domain-To observe loudness of signal	ZCR: Zero-crossing rate	Time domain-To observe high frequency contents
$S-BW$: Spectral Bandwidth	Frequency domain-perceived timbre of the sound, estimate vowel sound	$S-CENT$: Spectral Centroid	Frequency domain-estimate brightness of sound
$S-ROLL$: Spectral Roll-off	Frequency domain-estimate skewness of energy	STE: Short time energy	Distinguishes vibration signals from non -vibration signals
$Formant1-5$: Formant frequency	Frequency peaks in the spectrum which have a high degree of energy	$FB1-FB12$: Filter Banks	Frequency domain-Bandpass filters that separates the input signal into multiple components

4.3 Aerosol Regressor

Figure 5 provides a simplified representation of all ML-prediction models for aerosol estimation. It relies on five inputs from previous components: *activity_type*, *activity_loudness*, *mask_presence*, *electronic_voice* and *active_sources*. We build a regression model to estimate the rate of aerosol (*aerosol_val*) generated from two inputs, *activity_type*, and *activity_loudness*, using the Random forest regression model with a maximum depth of 5. Based on the *activity_type* (e.g., talk), the model is trained on the decibel level of speech audio data as a feature, following the findings in prior studies that the activity amplitude linearly increases aerosol generation [10]. This model is trained and tested using data recorded from the cleanroom study as described in Section 6.1.1 for two reasons. First, the cleanroom allows atmospheric particles to be controlled using HEPA or ULPA filters [44]. Second, given the controlled environment, measuring exhaled aerosol particles can be achieved using highly specialized lab equipment such as a Condensation Particle Counter (CPC). On the other hand, predicting aerosol emissions from detected human activities in a practical, real-world setting is naturally challenged by other atmospheric particles (unrelated to a person's respiratory activities) that act as confounding variables. Thus, our efforts mandated experiments in the cleanroom setting to develop an audio-based aerosol prediction mechanism in two parts.

4.3.1 ***Activity Type.*** Determining the occurrences of different human respiratory activities can be viewed as a binary classification problem with many activity recognition models operating simultaneously. Doing so also allows us to develop an extensible platform that can employ other trained models to support additional activity

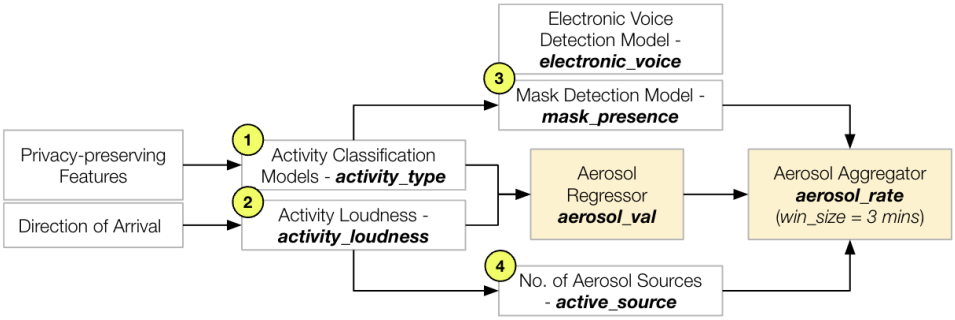


Fig. 5. Aerosol prediction pipeline.

types, which are not implemented but remain relevant in this work (e.g., singing and laughing). All classification models are built on the Random Forest algorithm as it can effectively handle imbalanced datasets. Models are trained extensively on publicly available datasets and tested on data collected from our user studies (we provide these details in Section 6). While public datasets consist of raw audio signals, all models are built on the small set of non-reconstructible features listed in Table 2, simultaneously allowing performance comparison between privacy-preserving and non-privacy models (i.e., trained using full raw audio spectrum).

Each activity model outputs an *activity_type* outcome of 1 or 0 (e.g., [1-talking, 0-not talking], [1-coughing, 0-not coughing], [1-sneezing, 0-not sneezing]), together with its prediction probability. In reality, it is plausible for a person to sneeze while someone else is talking. This component of activity classification will accordingly generate a list of activities detected to occur at the same time at 500 ms intervals (i.e., [talk, sneeze, cough, talk]).

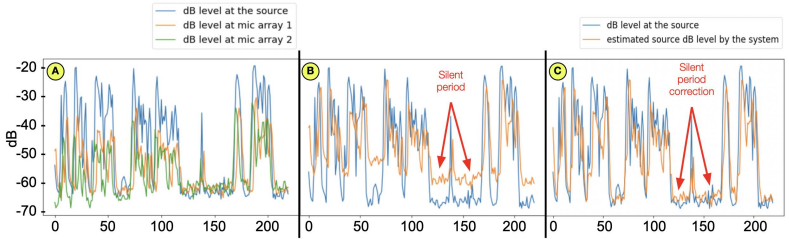


Fig. 6. a) Original dB level at Mic array 1, Mic array 2 and reference mic b) After Source dB level estimation using distance c)After silent-period based correction. blowup

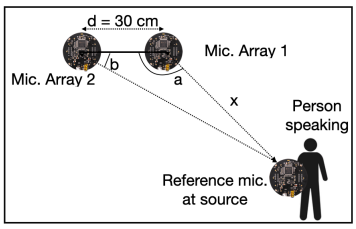


Fig. 7. Distance from the source to the sensor using trigonometry, solving for activity loudness.

4.3.2 ***Activity Loudness.*** After determining *activity_type*, our system estimates *activity_loudness* by calculating the distance between the received signal strength and the source signal strength. Figure 6 (a) shows the actual dB level at the source and the estimated dB level by the system. Our system must correct this difference to acquire accurate estimates of activity loudness. Given that speech amplitude linearly increases with aerosol generation [10], we can determine the loudness of the performed activity in two steps: Localizing the activity source and then calculating the signal strength of the activity source.

Step 1: Localizing Activity Source. Specifically, we use *dBFS* to measure sound intensity. For instance, *dBFS* for a 16-bit audio file is calculated as: $dBFS = 20 * log_{10}(abs(rms)/32768)$ because 16-bit signed has values between −32768 and +32767. This value, however, is not representative of the actual loudness generated at the

activity source because the received signal strength typically reduces with distance (between the speaker device placement and the human source). By employing two microphone arrays (see Figure 7), we can receive two direction of arrival (DOA) estimates, a and b, to calculate the distance estimation of the source. We place the microphone arrays at a fixed distance $d = 30cm$ apart and use a, b, and d to calculate the distance between the source of activity and microphone array 1, x. We find the distance of activity from mic_array_1, x, by iteratively solving the trigonometry equations for different angles of a and b.

Step 2: Signal Strength of Activity Source. Next, we use the estimated distance to calculate the actual decibel level at the source, $activity\_loudness$. Where signal strength at a far distance is given by: $L2 = L1 - |20 * log_{10}\frac{r_1}{r_2}|$, where $L1$ is the sound level at a distance, $r1$, from the source and $L2$ is the sound level at the farther distance, $r2$, from the source ($r2 > r1$) [45]. In this case, we know the signal strength at the farther distance $r2$, and we need to find the signal strength at the closer distance $r1$. Thus, we use the formula: $L1 = L2 + |20 * log_{10}\frac{r_1}{r_2}|$. For simplicity, we assume $r1 = 30cm$; that is, we find the signal strength at 30 cm away from the source (which would otherwise be infinity if $r1 = 0$).

4.4 Aerosol Aggregator

The aerosol aggregator is based on cumulative gains of $aerosol\_val$ at per second frequency over a sliding window of per minute interval. It is also impacted by two factors, (the number of) $active\_source$, and $mask\_presence$. Our sliding window parameter, $win\_size = 3$, builds on prior reporting that found aerosol droplets from speech remain active in the air between 8 to 14 minutes in a stagnant environment [46]. Note that win_size is an adjustable model parameter that can be changed based on environmental settings. For example, should an indoor space have a high ventilation rate, win_size can be altered to a lower value, denoting shorter intervals of accumulated aerosol droplets in the air.

4.4.1 ***Active Sources.*** As per Figure 7, the point at which an activity source is detected, defined by the DoA and estimated distance from mic_array_1, not only allows us to localize audio sources as part of activity loudness, it is valuable for estimating the number of "active" human sources engaged in simultaneously detected activities.

Using the estimated distance from mic_array_1, we retain a tolerance range of $\pm\epsilon$, as it is conceivable for the source to move around the detected space over time. At present, we streamlined our system pipeline to record a time-out window of $T_{out} = 3min$, where an aerosol source will drop off from the list should no new activity from their location be detected. When an activity is detected at a different and unmarked point, we assume this source is new and add it to the occupant list. An occupant can move from one place to another beyond the tolerance range of $\pm\epsilon$; thus, the system will record it as multiple occupants. This technique suffers from some inaccuracies, namely, undercounting sources that are very close together.

4.4.2 ***Electronic Voice Detection*** . Once speech activity is detected, we need to ensure that it does not come from electronic speakers (e.g., Zoom call speakers), as only human speech will contribute towards aerosol emission. To do so, we build on the techniques proposed by Blue et al. [47] to differentiate between the speech generated by humans and electronic speakers. This technique is based on the intuition that significant low frequency signals (sub-bass over-excitation) that are outside of the range of human voices but inherent to the design of modern speakers. This happens because audio created by an electronic speaker will have more energy in the sub-bass region due to the resonance of the enclosure. Our solution uses the energy balance metric defined as the ratio of energy in the sub-bass region (20-80 Hz) to the energy in the total evaluated region (20-250 Hz) as a differentiator between the two sources (sub-bass over-excitation). We calculate the energy balance metric by taking FFT (nfft=2048) of the 500 ms audio segment and extracting the energy in the desired frequency ranges. Figure 8 shows the energy balance metric values for an all-in-one PC speaker and a human speaker, both saying

the same phrase. This data is taken from the voice liveness detection dataset ASVspoof 2017 [48]. It can be observed that the energy balance metric values are significantly different for human and electronic speakers. Hence, we use a threshold-based binary classifier for electronic voice identification.[1] We use this threshold value to calculate the p-values for both the distributions and suppress electronic voices when estimating aerosols.

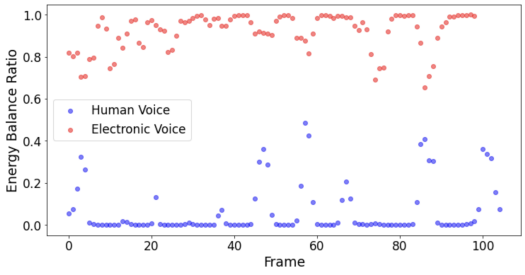


Fig. 8. Energy balance metric values for electronic and human speaker

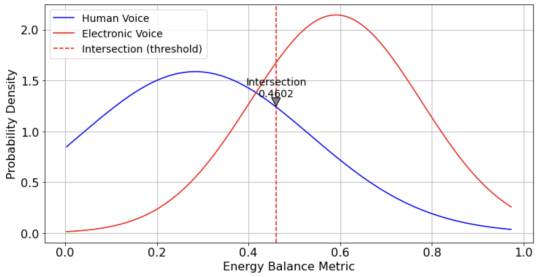


Fig. 9. Finding decision threshold to distinguish between genuine and spoof voice

4.4.3 ***Mask Presence.*** Indeed, surgical masks and KN95 respirators have reportedly reduced outward aerosol particle emission rates by 90% and 74% on average during speaking and coughing, respectively, compared to wearing no mask [12]. To this end, *mask_presence* is an essential parameter for our application to reduce aerosol estimation. Unlike prior work [14], we build a model for mask presence using *only* privacy-preserving features and a Random Forest classifier. Our model is trained extensively using our data and Mask Augsburg Speech Corpus (MASC) dataset [14] from ComParE Challenge 2020 that includes audio data recorded from 32 speakers wearing a Sentinex Lite surgical mask. We use the same training, development, and test corpus definitions as the challenge. While the training dataset consists of full spectrum audio, our model is built on the small set of non-reconstructible features listed in Table 2, simultaneously allowing performance comparison between privacy-preserving and non-privacy models.

For each detected activity in the list of activity output, the system runs the mask detection model, with its respective outcome of 1 or 0 (e.g., [1-mask, 0-no mask]). Accordingly, the result from this component will be used to correct the rate of aerosol estimation. Note, however, that our detection mechanism does not determine the kinds of masks being used by a user.

4.5 Transmission Risk Assessment

Aerosol droplets from humans typically dissipate within 30 minutes [49] and can remain suspended in the air for some time. For the suspended aerosol resulting from speaking, removal by gravity started only after 20 minutes [46, 50]. CDC guidelines defined close temporal proximity to be within 15 minutes [26]. Following these findings, we empirically capped the transmission risk window threshold, $trans_{win} = 15mins$. Further, we apply an upper bound on the product of the aerosol rate and an occupant's time in an enclosed space [31], a heuristic adapted from Bazant *et al.* in limiting indoor airborne transmission. We characterize these rates of aerosol generated during human activities as *low*, *medium*, and *high* depending on the activities detected, their loudness, the presence of mask, and the number of active sources altogether.

At this point, it is essential to highlight that the *AeroSense* is intended to work with standalone air quality sensors because estimating the risk of airborne transmission in the real world requires considering both *the*

[1]We decide the threshold of this energy-balanced metric using statistical analysis on the data recorded from humans and multiple electronic speakers using the ReSpeaker mic array. To do so, we extract energy balance metrics for electronic and human speakers and compute the mean and standard deviation of both groups. Next, assuming a normal distribution, we calculate the z-score at which electronic and human distributions intersect to find the threshold as shown in Figure 9.

rate of aerosol generation and *the rate of aerosol dissipation* factors, as motivated in Section 2.3. The latter, which includes ventilation rate and air filtration standards, is a set of characteristics measurable by existing air quality sensors (e.g., Netatmo for CO_2 [51] and *FlowSense* for ventilation rate [17]), however, the former is a set of key parameters our system aims to provide. With the ability of our system to determine *activity_type*, *activity_loudness*, *mask_presence*, *active_source*, and now, the rate of aerosol generated, *aerosol_rate*, these results can be logged to fulfill a complete risk assessment tool.

5 *AEROSENSE* IMPLEMENTATION

We implemented a prototype of *AeroSense* as shown in Figure 10, which we discuss below. The source code and datasets of *AeroSense* are publicly available at the github repository https://github.com/umassos/AeroSense.

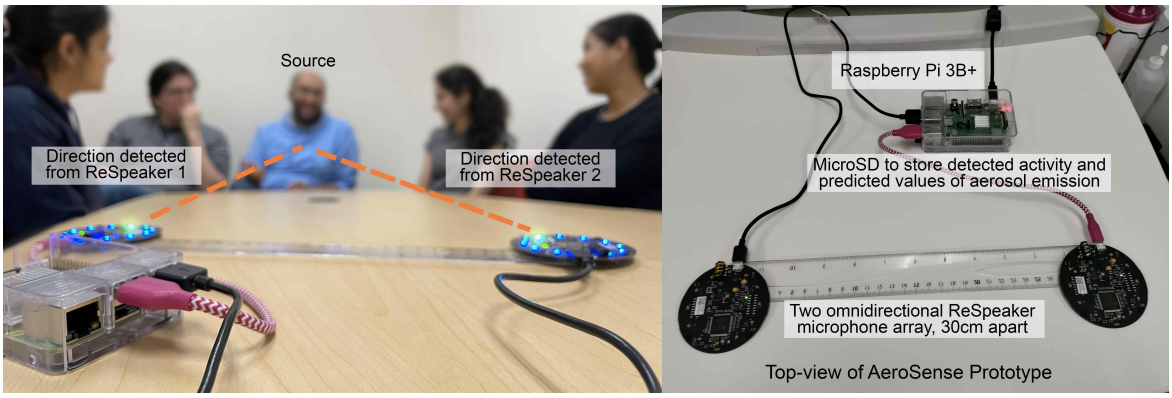


Fig. 10. *AeroSense* prototype using two omnidirectional speakers and Raspberry Pi 3B+.

5.1 Hardware Implementation

We used two ReSpeaker Microphone Array v2.0, placed 30 cm apart [52]. Each Microphone Array consists of four long-range microphones, recording samples at a 16 kHz sampling rate, and has a built-in capability of direction-of-arrival (DoA) estimation algorithms. We connected the two microphone arrays using a Raspberry Pi 3B+ and ran all feature extraction and model prediction components locally.

5.2 Software Implementation

AeroSense extracts privacy-preserving features every 500 milliseconds and the direction of arrival estimates from both microphone arrays at 100 ms.

Feature Extraction and Selection: We used Python to write the device driver layer script that extracts time and frequency domain audio features from the raw audio signal, specifically librosa[53], numpy, signal, and scipy [54]. For extracting DoA from microphone arrays, we use the built-in ReSpeaker driver code for the microphone array [55]. While training these ML models, we remove highly correlated features greater than 0.85. After this, we use impurity-based feature importance to select features based on high Gini importance.

Activity Classification: Next, we used the privacy features and passed them to speech, cough, and sneeze classifiers (trained using sklearn [56] models) individually to predict activity. If the activity model predicts 1, we add the corresponding activity to the list of detected activities, *activity_type*. If no activity is detected, *activity_type* remains empty, and we do nothing. If activity is detected, then for all the activities in *activity_type*,

we perform four main steps: a) activity localization and loudness estimation, b) update active sources list, c) predict aerosols, and d) detect mask.

a) Activity localization and loudness estimation: If the activity is detected at timestamp t, we find the distance of the activity using the DoA values at timestamp t by using trigonometric calculations. We use this distance to calculate $activity\_loudness$ if it is not a silent period. In the presence of a silent period, we do not perform dB level extrapolation. We use $silence\_threshold$ as -55 dB.

b) Update active sources list: We updated the number of active aerosol sources $active\_sources$ list every time an activity is detected, which is the list of aerosol emitters, their locations, and the last timestamp when they were active. We use the angle threshold and distance thresholds ϵ and timeout threshold T_{out} for this. If the activity location lies within the thresholds of any existing occupant, we update the location of the current occupant. However, if the location of an activity does not lie within the thresholds of any existing occupant in the $active\_sources$ list, we add a new occupant with its location and timestamp. Additionally, we check for timeout; if any occupant in the $active\_sources$ list has not performed an activity since T_{out} time, remove the occupant from the list. We set these parameters as $T_{out} = 3min$ and angle threshold $\epsilon = 20$, with a distance threshold of 80 cm.

c) Predict aerosols: Next, we predict aerosols using $activity\_type$, $activity\_loudness$, $active\_sources$, and $mask\_presence$. If the $activity\_type$ is cough/sneeze, we use the aerosol/cough and aerosol/sneeze numbers to get $aerosol\_val$. If the activity is speech, we call the regression model that takes $activity\_loudness$ as input and predicts the aerosol emissions due to speech $aerosol\_val$ at second-level. This regression model is trained using a Random Forest model.

d) Detect Voice Liveness: We use the energy balance metric to classify between electronic and human speaker $electronic\_voice$. The electronic voice detection model uses a threshold-based classifier, with a decision threshold of 0.4. If a human voice is detected, $electronic\_voice$ is set to 0, else 1. If it is 1, we skip the following step of aerosol estimation and assign $aerosol\_val$ to 0 for that segment.

e) Detect Mask: We also use the privacy features as input to the mask classifier model to detect $mask\_presence$. The mask classifier is trained using a gradient-boosting classifier model. If a mask is detected, $mask\_presence$ is set to 1. If a mask is present, we multiply the aerosol emissions, $aerosol\_val$, by 0.4 as we observed the aerosol reduction by 60% due to mask in our experiments (Section 7.1.6).

Lastly, we aggregate aerosols by adding the aerosols of the current window and the previous win_size window. We use $win\_size = 3min$ in this paper.

5.3 Deployment Considerations

AeroSense can be deployed in two configurations, standalone or in conjunction with an existing Building Management System (BMS). In the former case, *AeroSense* devices deployed in various building rooms communicate with a central server to estimate transmission risk in each instrumented room and notify users as needed. In the latter case, each *AeroSense* device communicates with the BMS, which also integrates it with the building's ventilation system. This can enable additional measures, such as dynamically increasing ventilation in spaces with higher transmission risk. As designed, *AeroSense* is cost-effective, with the bill of materials amounting to $280 per instrumented room. In contrast, traditional particle counters come at a significantly higher cost, exceeding $16,000.00 per device, and pose logistical challenges in deployment due to their size and weight. *AeroSense* is a more practical solution, offering scalability and ease of maintenance. *AeroSense* ś maintenance overheads are similar to conventional IoT sensing systems. As a powered system, there are no battery replacement overheads, but there is a calibration phase to estimate the silent period threshold and other parameters. Recalibration may be necessary if the physical configuration of the room changes significantly, but this is typically infrequent.

6 EXPERIMENTAL METHODOLOGY AND DATASET DESCRIPTIONS

This section describes our data collection methodology, additional public datasets used in our evaluation as well as our real-world and in-the-wild deployments.

Data Ethics. We received approval from our university's Institutional Review Board (IRB) to deploy *AeroSense* device prototype and conduct user studies under realistic settings. Participant data for all real-world experiments was collected with user consent. We only gathered privacy-preserving audio features, with no raw audio, to safeguard privacy.

6.1 Data Collection Methodology

6.1.1 ***Cleanroom User Study.*** The first user study took place in a cleanroom setting with a HEPA-filtered laminar flow on the ceiling to minimize background particle concentration, allowing us to collect actual values of aerosol emissions using advanced microbiological technology while performing speech activities. The ability to predict aerosol generated and verify against accurate aerosol measurements are typically performed in cleanroom settings for two reasons. First, the cleanroom allows atmospheric particles to be controlled using HEPA or ULPA filters [44]. Second, given the controlled environment, measuring exhaled aerosol particles can be achieved using highly specialized lab equipment such as a Condensation Particle Counter (CPC). On the other hand, predicting aerosol emissions from detected human activities in a practical, real-world setting is naturally challenged by other atmospheric particles (unrelated to a person's respiratory activities) that act as confounding variables. Thus, our efforts mandated experiments in the cleanroom setting to develop an audio-based aerosol prediction mechanism in two parts.

Table 3. Experimental setup for component-specific data collection and evaluation.

	Cleanroom	Activity Detection	Activity Loudness	Mask Detection	Electronic Voice
Duration	10 hrs collected	20 hrs opensource	2 hrs collected	10 hrs opensource and 5 hrs collected	5 hrs collected
Location	Cleanroom	N.A., Google	300 ft^2 lab	MASC, 300 ft^2 lab	300 ft^2 lab
No. of users	1F	N.A.	1M, 1F	3M, 2F	3M, 2F
Ground truth	CPC 3775	Annotation files	Participants wearing ReSpeaker Mic	Different files for mask and no-mask	Different files for electronic and human
Data Usage	80% train 20% test, 5-fold CV	80% train 20% test, 5-fold CV	100% test	80% train 20% test, 5-fold CV	100% test
Dataset	.wav files, aerosol csv files	.wav files, .TextGrid files	csv files (containing dB level)	audio feature csv files	audio feature csv files
Activities	Speaking at different loudness, coughing, sneezing, breathing	Speech, Cough, Sneeze, Ref. Table 4	Reading	Reading	Reading

As illustrated in Figure 11A, a consented participant would use a facepiece respirator while reading scripted texts in different decibel ranges in the room. This respirator is directly connected to a CPC Figure 11B, which outputs aerosol particle concentration measured in p/cc. We gathered ground truth aerosol count using a Condensation Particle Counter 3775, a general-purpose particle counter device that detects airborne particles down to $4nm$. We set the flow rate of the particle counter to 1.5 L/min, and it records data every 1 second. This setup, however, does not allow us to deploy *AeroSense* because it requires our participant to put on a facepiece respirator Figure 11C, which will inherently affect the real-time extraction of feature samples. As a workaround, we recorded raw

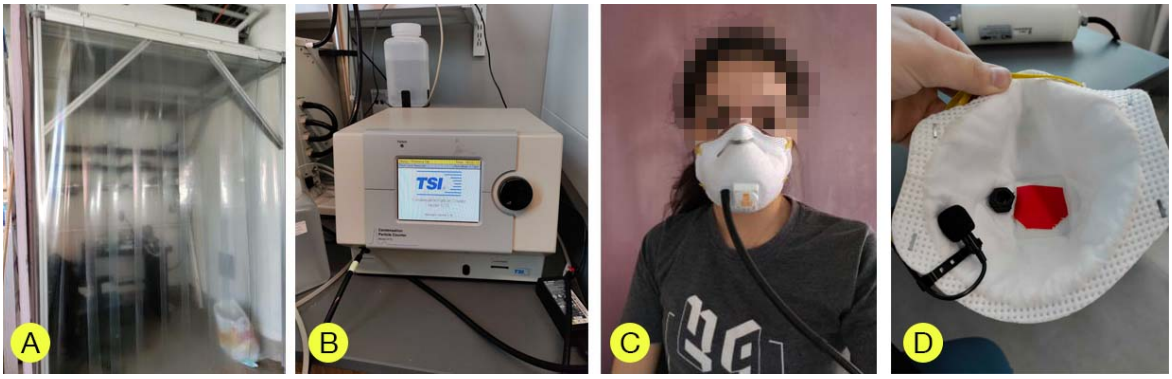


Fig. 11. Cleanroom experiment to develop *AeroSense*.

audio signals from a microphone fixed inside the facepiece Figure 11D. This effort amounted to ≈ 10 hours of speech data with and without masks, respectively, with their aerosol particle emissions, over seven days.

One participant carried out scripted activities of reading texts in different loudness, with and without wearing a mask, amounting to over 10 hours of audio features and aerosol values. To record most of the aerosols emitting from a participant's mouth due to the activity, we connected a duct to an N95 mask (with a sealed filter) and the other end of the duct to the sensor. This way, we could record only the particles emitted from the user's nose and mouth. We placed a microphone inside the mask to record the audio data and the aerosols (See Figure 11D).

6.1.2 ***Activity Detection***. We utilized the Google AudioSet, an open-source dataset offering an extensive collection of annotated 10-second sound clips from YouTube [57]. From this collection, we selected clips containing cough, sneeze, speech, silence, sniffle, sneeze, gasp, breathe, throat-clearning, hiccup, vomit, burp, wheeze, snore, and variation of indoor background noises, amounting to approximately 20 hours of sounds as shown in Table 4. While clips related to cough and sneeze make up 17.16% of this dataset (≈ 3.4 hours of data), the assessment of all categories was reported to be high quality. We used this dataset to train a combination of privacy-preserving activity classification models. For each model, the activity class is labelled as 1, and the 0 class includes all the other activities. For instance, the cough classifier has cough samples labeled as 1 and all the other activities (e.g., speech, sneezing, silence, sniffling) labeled as negative class. For the cough classifier, we include cough-like sounds in the negative class to test the robustness of the model against false positives.

Table 4. Duration of various activities in the Google AudioSet. Others include gasping, breathing, snoring, hiccuping.

Label	Percentage (%)
Speech	39.95
Background Noises	37.00
Cough	15.09
Silence	2.69
Sneeze	2.07
Sniffle	1.44
Others	1.76

Table 5. Duration of mask and no-mask data

Label	Duration (hrs)
No Mask (MASC data)	4.96 hrs
No Mask (our data)	2.5 hrs
Surgical Mask (MASC data)	5.36 hrs
N95 mask (our data)	2.5 hrs

6.1.3 ***Mask Detection.*** In training our model, we used our own recorded dataset consisting of only non-reconstructible audio features collected from 5 consented participants(3 males and 2 females, aged 23 to 28 years) wearing an N95 mask. This dataset amounted to ≈ 5 hours of sounds of people talking, coughing, and sneezing, with and without a mask (see Table 5). In addition to this, we use the Mask Augsburg Speech Corpus (MASC) dataset from ComParE Challenge 2020 that includes audio data recorded from 32 German native speakers (16 females, 16 males, age from 20 to 41 years, mean age 25.6 years, standard deviation 4.5 years), wearing a Sentinex Lite surgical. We use the same training, development, and test corpus definitions as the challenge. The audio was sampled at a rate of 48 kHz with 24-bit, downsampled, and converted to 16 kHz and mono/16 bit; the total duration is 10 h 9 min 14 sec. The participants performed different tasks without masks. While wearing the mask: They answered some questions, read words known for their usage in medical operation rooms, drew a picture and talked about it, and described pictures, e.g., sports activities, families, kids, food, or locations. We used Train/Dev/Test sets defined in the challenge for the MASC dataset, with 12 speakers in Train and 10 in Dev and Test each.

6.1.4 ***Activity Loudness Estimation.*** At different times, two participants (1M/1F) were directed to read a scripted Wikipedia text at various loudness and distance, ranging from 50cm to 3m, away from the prototype. The participants were wearing another ReSpeaker mic to record the ground truth values of loudness in decibels.

6.1.5 ***Electronic Vs Human Voice.*** For electronic and human voice differentiation, we recorded data from 5 consented participants and 5 electronic speakers: a MacBook Pro, Alexa, TCL TV, OnePlus Nord N10 smartphone, and iPhone 15. We recorded data for 30 minutes from each electronic speaker and participant (a total of 5 hours). We asked participants to read different Wikipedia texts and play podcasts from different speakers.

6.2 *AeroSense* Deployment

To evaluate the efficacy of *AeroSense*, we deployed our prototype in different types of rooms in our academic building. We used it to conduct user studies in real-world and long-term in-the-wild settings.

6.2.1 ***Real-world Studies.*** We performed a number of user studies using our prototype in four different types of indoor spaces, where we collected over 25 hours of privacy-preserving audio data (see Table 6).

- *Office room (lower-risk):* 2-3 occupants engaged in a weekly meeting lasting for 5 hours in an office.
- *Large conference room (higher-risk):* 5-20 occupants engaged in a weekly meeting for 15 hours in a seminar room. One instance is depicted in Figure 13, where occupants were engaged in a group meeting. In this scenario, we also record zoom calls data as some questions were raised via Zoom.
- *Small Conference Room:* 5-8 occupants engaged in a discussion for 3 hours.
- *Classroom:* 35 students engaged in a discussion session in a classroom scenario for 2 hours.

The lower risk and higher risk scenarios noted above are designed to show the significance of aerosol estimation in a realistic setting as follows:

- *Lower Risk Activity:* Two participants (1M,1F), seated approximately 2.5 m apart, conversed for 30 minutes at approximately -25 dB average loudness in office room (see Figure 12)
- *Higher Risk Activity:* Five participants (3M/2F), seated approximately 60-240cm apart, conversed for 1 hour at approximately -20 dB average loudness (see Figures 12 and 13).

Data Description: In all cases, our prototype records two main files: a) location file containing the angle of arrivals at the two microphone arrays, distance with timestamps, and b) audio features file containing privacy-preserving features with timestamps. A human observer collects the ground truth by taking logs of activities and active sources at per-minute granularity.

Table 6. Experimental setup for system-wide data collection and evaluation.

	Office room	Large conference room	Small conference	Classroom	Long-term deployment
Duration	5 hrs	15 hrs	3 hrs	2 hrs	~160 hrs
No. of users	2-3	5-20	5-8	35	0-8
Activities	Weekly meetings	Lab meetings	Meetings	Discussion session	People working, chatting

Ground truth: We rely on a human observer to collect ground truth information at a per-minute level. The observer records active sources and activity information with time. The location file is used to verify active sources, and the privacy-preserving audio features file is used to detect the activity and its loudness. We use the distance estimation to correct the activity loudness.

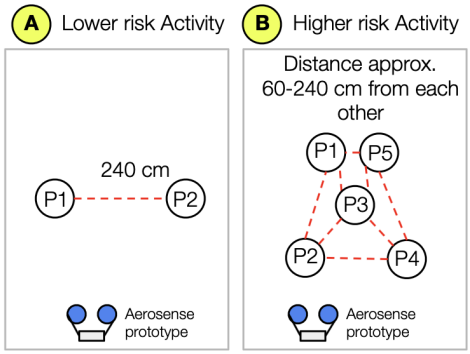


Fig. 12. Participants in scripted scenarios of our real-world experiments.

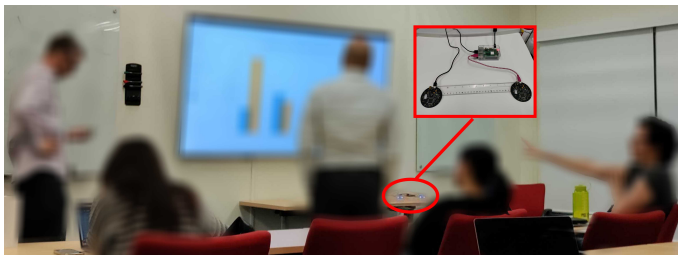


Fig. 13. Real-world study of a 5-person (unmasked) meeting in large conference room, simulating a higher-risk scenario.

6.2.2 ***Long-term In-the-wild Deployment.*** Our in-the-wild deployment was done in a medium-sized lab setting over the course of multiple weeks. During this time, we manually collected ground truth from 9 a.m. to 5 p.m. over 3 weeks, which resulted more than 160 hours of deployment data. Like before, the data consists of location files, feature files, and ground truth logs.

7 SYSTEM EVALUATION

In this section, we evaluate the accuracy of our *AeroSense* in estimating aerosol generated from human activities. Further, we assess the components responsible for producing critical inputs to our aerosol prediction mechanism.

7.1 Results

Evaluation Metric: Our classification models prioritize recall to accurately determine an activity type occurring, whereas high precision implies that an activity of interest did not occur. F1-score is the weighted average of precision and recall. Our aerosol emissions predictions are evaluated based on mean-squared error (MSE) for the regression model, where a lower MSE closer to 0 is best.

7.1.1 ***Efficacy of Audio-based Aerosol Prediction.*** The crux of our work lies in *AeroSense*'s ability to predict aerosol generation from detecting different human activities using audio features. As a recap, upon detecting a specific human activity, in this case, $activity\_type = talking$, *AeroSense* employs a regression model for speech to estimate the amounts of aerosol generated from this event. The model is built using a single feature, dB level, representing $activity\_loudness$.

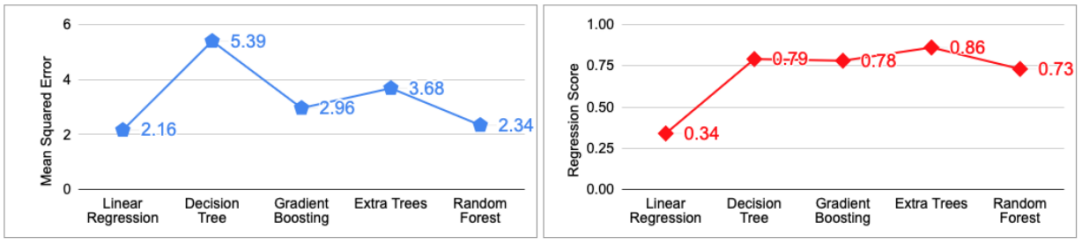
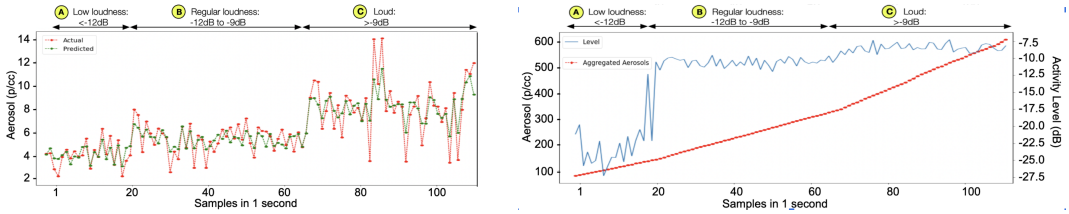


Fig. 14. Aerosol Regressor model performance upon detecting speech as an activity.



(a) Predicted aerosol amounts using a Random Forest regression model.

(b) Increasing rate of aerosol over time, assuming no loss of particles ($win\_size = 3mins$)

Fig. 15. Aerosol Prediction Efficacy

Figure 14 summarizes the model performance using different regression techniques to predict aerosol amounts, *aerosol_val*, upon detecting an activity. In validating the models, we split the *Cleanroom* dataset into 80% for training and 20% for testing our aerosol prediction model. Where values closer to zero are better, our experiments found the Random Forest algorithm performing best. Although the mean squared error of 2.34 is not the lowest, the overall regression is statistically significant (R^2 = 0.73, p = < .001) compared to employing Linear Regression.

Next, we used our regression model built on Random Forest to aggregate the rate of aerosol generated at different speech loudness of low, regular, and loud (note: the decibel levels categorizing loudness in *Cleanroom* is different from real-world studies due to the unique setup of a microphone closer to mouth (see Figure 11). We plotted the *aerosol_rate*, measured in p/cc, at 1-second intervals, as per Figure 15a (top). Further in Figure 15a (bottom), we charted aerosol accumulation, with no loss of aerosol particles assumed within a 3-minute window. As expected, the rate of aerosol will increase over time if aerosol dissipation is not taken into account. The loss of particles will depend on the number of surfaces, temperature, humidity, ventilation, and filtration. For the scope of this work, we only focus on estimating aerosol emissions. An insightful observation is that aerosol rates increase with activity loudness.

Key Takeaway: The main contribution of our work lies in exploring an audio-based prediction mechanism to estimate the rate of aerosol from human-generated activities. Using a Random Forest regression model, the model takes *activity_loudness* (i.e., dB level) as a feature upon determining an *activity_type*. It then aggregates the amount of aerosol over a 3-minute window, where no particle loss is assumed during this time. However, this threshold is adjustable based on the ventilation rate in the dedicated space. Our model can only be validated in a *Cleanroom* using a CPC to control for aerosol particles that may not be attributed to human-generated activities.

7.1.2 ***Privacy-Preserving Activity Classification Models.*** Indeed, the accuracy of our aerosol prediction mechanism above is impacted by the subcomponents responsible for two critical inputs: *activity_type* and *activity_loudness*. Here, we systematically present the results of our activity classification models. Using 100% of the *OpenSource* dataset for training and testing on our real-world dataset, we compare performances between models using an unfiltered audio spectrogram and privacy-preserving features.

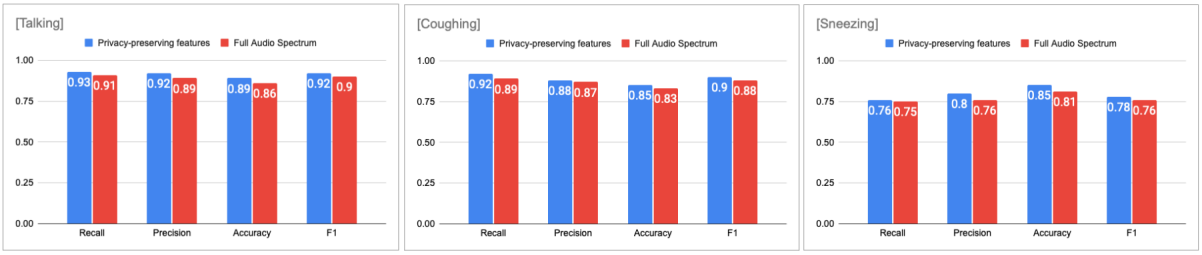


Fig. 16. Model performance of activity classification models using privacy-preserving features and full audio spectrogram.

Figure 16 charts the recall, precision, accuracy, and F1-score of three human respiratory activities of interest: talking, coughing, and sneezing. Similarly, using a Random Forest algorithm but now solving for a binary classification problem, the models trained using only privacy-preserving features achieve better performance on all metrics than training a model with a full audio spectrogram. Where recall is prioritized, our models achieved 93% for speech, 92% for cough, and 76% for sneeze, respectively. Both model types, however, did not yield significant differences. We compare our accuracy with the state-of-the-art cough classification model CoughBuddy [58] and achieve a comparable accuracy of 86.56%, recall of 92.07%, precision of 89.28%, and F1 score of 90.65%. We use audio features proposed in this work on our dataset, and trained a RF classifier on it. However, this work does not focus on extracting privacy-preserving audio features and uses features like MFCC, which contain sensitive information.

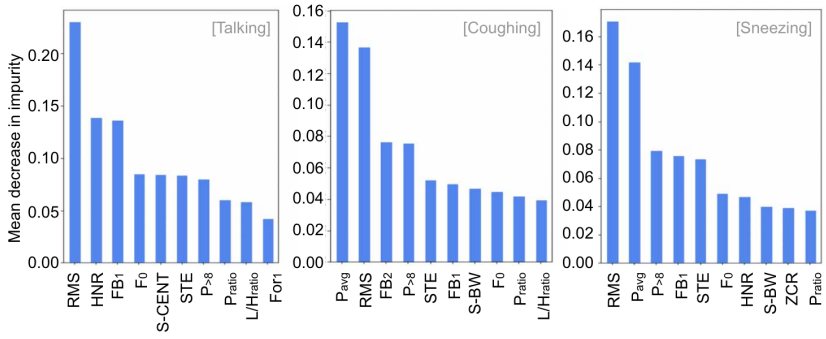


Fig. 17. Feature Importance for privacy-preserving activity classifiers.

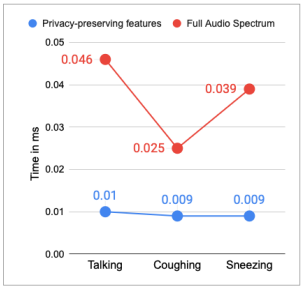


Fig. 18. Prediction latency comparing non/privacy-preserving activity models.

Figure 17 ranks ten useful features by gini importance. The results on these features align with findings from prior work, and accordingly, our discussion in Section 4.2, where we conjecture them to be useful for detecting users talking, coughing, and sneezing. For example, P_{avg} (mean power), RMS (root mean square energy), $P_{>8}$ (power above 8 kHz), FB_1 (Filter bank with center frequency 308.92 Hz), and STE (short time energy) are equally important for detecting cough and sneeze. Cough and sneeze are very similar in nature, where air bursts suddenly through the lungs with force, producing loud but brief sounds. Notably, ZCR (zero-crossing rate) distinguishes

sneezing from coughs. On the other hand, a unique feature for detecting a person talking is the F_0 (base frequency), which is a fundamental acoustic cue for talkers [40].

Constraining our detection algorithms to only a small set of audio features helps preserve privacy and significantly reduces the model latency in predicting an activity. As shown in Figure 18, determining whether a person talks, coughs, or sneezes can be accomplished in under half the time required using the entire audio spectrogram.

Key Takeaway: It should be noted that activity classification of detecting a person talking, coughing, or sneezing is not new. Prior work has shown several methods of preserving privacy in these situations, including selective sampling for non-speech periods and suppressing the full audio spectrogram. Our work explores combining these key non-reconstructible audio features to accurately match the performance of activity classification with a privacy-invasive method. Our experiments yield insignificant differences in accuracy between both model types and lower latency. *AeroSense* requires and favors such models, as the system adopts a privacy-first as its key design and runs on the edge.

7.1.3 ***Activity Loudness Estimation.*** We have made the case that the dB level representing an activity's loudness will impact aerosol estimation. Indeed, our aerosol regression model (see Figure 14) requires a single dB level as its model input. Recall the problem motivated in Figure 6 illustrates significant differences between the dB level at the source compared to the dB level at the two microphone arrays of the system, which reduces with increasing distance. Thus, our system requires the capability of correcting activity loudness. We evaluate the accuracy of our system in determining the actual activity loudness at the source by finding the prominent peaks in the dB level signals and then calculating the average of the peaks. We find peaks in the signal to avoid silent periods in evaluating loudness estimation. We compute this average for the source signal, received signal at microphone array 1, and estimated signal (source signal estimated by the system).

Table 7 summarizes our results from conducting the activity loudness study, where participants were directed to talk at varying distances (between 50cm - 3m) and loudness (low: <-24dB, regular: -24dB to -18dB, loud: >-18dB). Overall, the system produces an average error of 7.74% in predicting source activity level. We observed increased errors in received signals as the distance between participants and the microphone source increased. Nonetheless, by applying the loudness estimation technique, these errors were reduced from ≈ 23% to 7% for 2m, ≈15% to 8% for 1m, and ≈9% to 7% for 50cm, on average. This result is anticipated as the received audio signal will be similar to the source when they are in close proximity.

Table 7. Activity loudness corrected from dB level extrapolation and silent-periods for one participant.

distance/dB		Source	Received	Received error	Estimated	Estimated error
50cm	Low	-30.93 dB	-34.32 dB	11.0 %	-29.88 dB	3.4 %
	Reg	-27.76 dB	-29.92 dB	7.8 %	-25.58 dB	7.9 %
	Loud	-28.56 dB	-31.06 dB	8.8 %	-25.49 dB	10.7 %
1m	Low	-33.41 dB	-37.1 dB	11 %	-30.94 dB	7.4 %
	Reg	-26.85 dB	-31.01 dB	15.5 %	-24.56 dB	8.5 %
	Loud	-24.92 dB	-29.76 dB	19.4 %	-22.74 dB	8.7 %
2m	Low	-31.4 dB	-38.92 dB	24 %	-29.81 dB	5.1 %
	Reg	-28.56 dB	-36.46 dB	27.7 %	-26.02 dB	8.9 %
	Loud	-26.49 dB	-31.31 dB	18.2 %	-28.91 dB	9.1 %

Key Takeaway: It is important that we achieve a highly accurate capability to estimate activity loudness. The consequence of its inaccuracies will affect the model performance of *AeroSense* in predicting aerosol amounts as the model is built on a single *RMS* feature, representing the loudness of detected activity.

7.1.4 ***Active Source***. Correcting for activity loudness from the received signal simultaneously allows our system to localize the active (human) source engaged in a detected activity, as motivated in Section 4.4.1. It is conceivable that an active source moves around a designated space within a tolerance range, $\pm\epsilon$. It is equally possible that an active source switches to inactivity over T_{out}. Estimation parameters are set at $T_{out} = 3min$ and angle threshold $\epsilon = 20$, with a distance threshold of 80 cm. Since our system only uses activity detection and localization to approximate active sources, it is essential to note that *active_source* does not equal occupancy size (i.e., there can be occupants in the room who did not engage in human respiratory-type activities).

While our aerosol regressor model predicts aerosol amounts on a second granularity, our ground truth observational logs for active sources were recorded in 1-minute intervals. Consequently, in evaluating the number of sources, we consider the maximum number of active sources for the entire 1-minute duration. Figures 19 and 20 (top) chart the number of active sources our system can accurately detect in one-minute samples for two experiments as part of our real-world study: lower and normal-risk activity. Figures 19 and 20 (bottom) correspondingly plot the predicted aerosol amounts based on activity loudness, as instructed to participants. Since participants engaged in an open conversation, our setup did not control who should speak and when.

Nonetheless, our estimation technique generally predicted the number of active sources every minute accurately. For example, as per Figure 20, the errors were attributed to participants sitting close to each other, leading to the underprediction of active sources. An additional observation is that if sources are moving while talking, the system will overpredict the active sources as it will keep sensing activities at new locations continuously before the timeout.

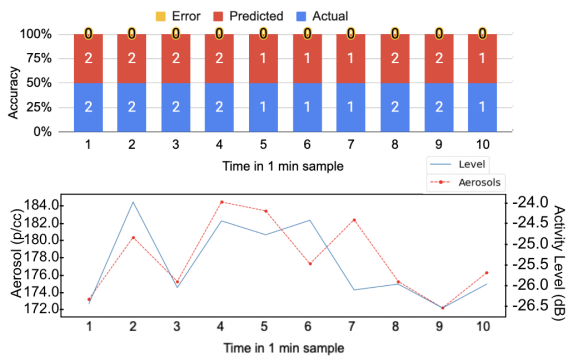


Fig. 19. (Top) Results for predicting active sources in lower-risk real-world study. (Bottom) Predicted aerosol amounts based on activity loudness.

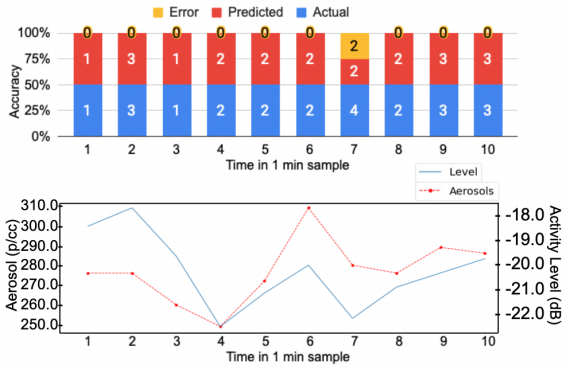


Fig. 20. (Top) Results for predicting active sources in normal-risk real-world study. (Bottom) Predicted aerosol amounts based on activity loudness.

Key Takeaway: Correcting activity loudness allows us to determine the distance between various sound sources, which sequentially helps us localize a source and determine the number of active sources engaged in detected activities. This approximation makes a reasonable assumption that active sources are not in close proximity and are not highly mobile to represent most everyday indoor activities. Since the approximation of active sources is based on an audio feature, this technique is inherently incapable of accounting for silent occupants.

7.1.5 ***Voice Liveness Detection***. As discussed in Section 6, we collected data with different smart speakers and human speakers to evaluate voice liveness detection.

We use True Positive Rate (TPR) at the optimal threshold as the metric to understand the performance of the classifier. TPR is the probability that a human voice score is less than or equal to the threshold. It can be

Table 8. Performance of voice liveness detection with different electronic speakers

Electronic Speaker	Optimal Decision Threshold	TPR (%)
Smartphone 1	0.4350	79.89
Smartphone 2	0.4601	80.54
TV	0.3943	74.10
Laptop	0.2769	52.09
Alexa	0.2553	50.06

observed from Table 8 that for TV and smartphones, we can achieve a threshold that can distinguish between electronic and human voices with 80% accuracy. However, for Laptops and Alexa (smart speaker), we cannot find a threshold that can distinguish between them because the energy balance metrics range for laptops and Alexa overlaps significantly with human speakers. To this end, we hypothesize that the energy balance metric-based technique is insufficient for advanced speakers and needs additional audio features to make it generalizable across all speakers.

7.1.6 ***Privacy-Preserving Mask Detection.*** The fourth and final critical input for our aerosol prediction mechanism is detecting mask presence among active sources. We collected data our data as well as used *OpenSource* MASC dataset. Accordingly, we conducted a leave-one-user-out cross-validation among 5 participants (see Section 6) and reported the average performance on our data. For the MASC dataset, we used Train/Dev/Test sets defined in the challenge, with 12 speakers in Train and 10 in Dev and Test each. We achieve an Unweighted Average Recall (UAR) score of 71.12% on the test set of MASC dataset (when trained and tested on MASC data) This is comparable to the state-of-the-art UAR of 71.8%, which is achieved by fusing the results of different feature-based learning like ComParE acoustic feature set and Bag-of-Audio-Words [59].

By using only non-reconstructible features, our results yield ≈ 75% recall and precision Random Forest algorithm (when trained and tested on our dataset). We hypothesize that the reason for the higher accuracy of our dataset is that it is recorded using a KN95 mask, which is thicker than the surgical mask used in MASC data.
When we combine MASC and our data and use it for training and testing, we achieve 71% recall and 72% precision as shown in Figure 21. Further, the results on feature importance, as per Figure 22, support prior reporting of LH_{ratio}, $P_{>8}$, and HNR to highly correlate with users wearing masks [41]. Our findings on feature importance shed some insights into incorporating mask detection into the system pipeline. Specifically, since the aerosol regression model depends on the loudness of activity (*RMS* as a feature), and conversely, for mask detection *RMS* is not the strongest predictor, we conjecture that despite a successful detection of mask among active sources, it will less likely impact the prediction of aerosol amounts.

The above finding helps inform a critical design decision of utilizing *mask_presence* to influence aerosol aggregation rather than directly altering aerosol prediction (see Figure 5). To better understand how *mask_presence* will impact aerosol amounts, we assess the reduction rate of aerosol particle counts measured directly using the CPC in *Cleanroom*. More specifically, Figure 23 charts the aerosol amounts in p/cc from a user talking without a mask (Segment A) and with a surgical mask (Segment B). With mask presence, we can expect the aerosol particles emitted from human activities to reduce by 60%. This result aligns with prior reports on the effects of mask-wearing to reduce aerosol emissions by 60%-90% [12]. Accordingly, in our implementation moving forward, we multiply the aerosol values by 0.4 if the presence of a mask is detected.

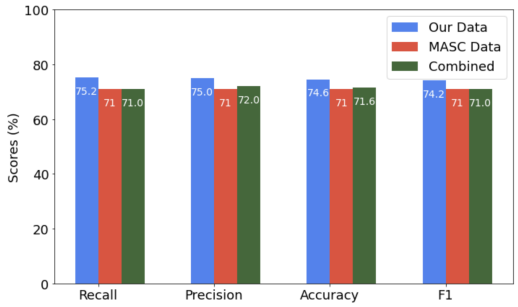


Fig. 21. Model performance of mask detection using privacy-preserving features.

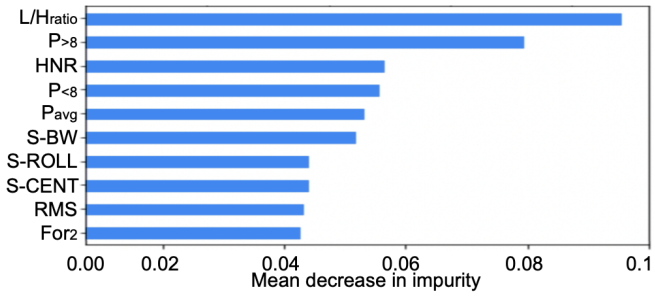


Fig. 22. Top 10 non-reconstructible audio features for mask detection.

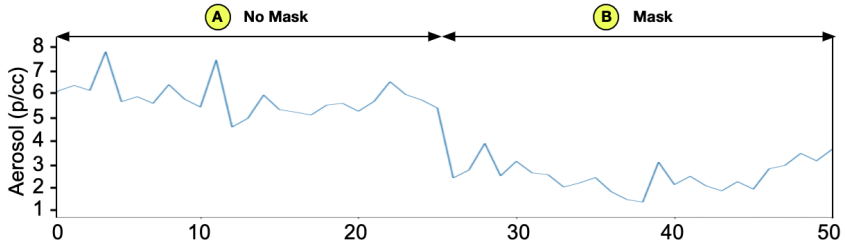


Fig. 23. Decreased aerosol amounts from talking with mask, measured using CPC.

Key Takeaway: Our work contributes to the investigation of detecting mask-wearing by using only privacy-preserving audio features. While a privacy-preserving mask detection model aligns with *AeroSense*'s key design choice, we found that the insignificance of the *RMS* audio feature (*activity_loudness*) does not directly affect the prediction of *aerosol_val* using the regression model. A reasonable workaround is to implement the reduction of aerosol amounts as a switch case where 60% of *aerosol_val* is reduced when *mask_presence* = 1.

7.1.7 ***AeroSense's Performance in Different Environments.*** As mentioned in Section 6, we extensively evaluate our system in different environments by deploying it for hundreds of hours. The deployed *AeroSense* prototype records privacy-preserving audio features and location files. For these deployments, we rely on a human observer to collect logs of activities and several active sources at a minute level. We compute minute-level accuracy for activity detection and active sources. To estimate activity detection accuracy at the minute level, we first find the accuracy of each one-minute segment and then take an average. For each minute segment, we have multiple predictions as our sliding window size is 500ms. To find activity detection accuracy at the one-minute segment, we find the maximum repeating label in that segment and see if that matches the actual label (logged by the observer). We calculate active sources at minute-level by using a similar technique (Section 7.1.4). For estimating average error in active sources, we compute the ratio of the sum of the absolute value of error in each minute segment to total segments.

Table 9 shows the average accuracy of activity detection and error in detecting active sources. It can be seen that the proposed system detects activity with a very high accuracy of >95%, whereas long-term deployment has the lowest accuracy of 95.8%. This is because more corner cases were observed in this deployment, like similar-sounding environment noises. In this scenario, the activity is almost always speech, and cough was observed only 7 times. The system correctly classified all these coughs. We observe a false positive rate (FPR) of

0.036-0.18% by the cough classifier in different environments. Sneeze was neither detected nor observed in these deployments. Our system shows a low average error in sensing active sources. The maximum error was observed in a laboratory environment. We hypothesize that this is caused by closely positioned desks in the laboratory.

Table 9. Performance of *AeroSense* in Different environments

Environment	Avg. Activity Detection Accuracy (%)	Avg. Active Sources Error (±)
Small Office Room	98.3	0.110
Large Conference Room	97.9	0.278
Small Conference Room	100	0.228
Classroom	97.7	0.189
Laboratory (Long term deployment)	95.8	0.45

7.1.8 ***AeroSense's Performance in Long term deployment.*** We conducted our experiments by deploying *AeroSense* in a Lab without controlling where participants are localized in the room and how loudly they speak. Figures 24 and 25 chart the predicted aerosol amounts minutes under two situations comparable to lower-risk and higher-risk activity that lasted for 30 minutes. Where a lower-risk activity consisted of two people talking in a meeting room and a higher-risk activity consisted of 8 people actively participating in a group meeting, we observed similar findings to our *RealWorld* experiment of the predicted *aerosol_val* corresponding with the captured activity loudness in these scenarios.

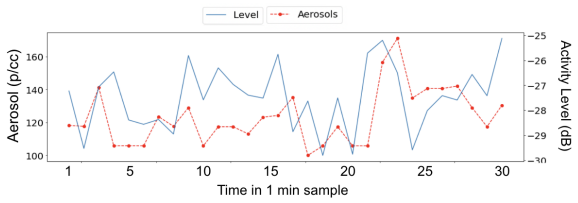


Fig. 24. Predicted aerosol amounts for lower-risk activity.

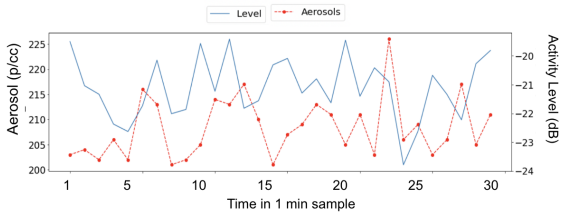


Fig. 25. Predicted aerosol amounts for higher-risk activity.

Figure 26 shows the rate of aerosol aggregated over a 3-minute window comparing these activities. We can observe more *aerosol_rate* in a higher-risk setting compared to a lower-risk one. Considering that our aerosol prediction mechanism operates on the same unit of measurement of a CPC at 1.5 L/min, a value of $1000p/cc$ is comparable to $1.5 \times 10^6 P/min$. This result demonstrates continuous speaking, especially at high volumes, can pose a considerable risk of indoor airborne transmission to occupants, as a significant fraction of speech particles can remain dispersed in the air for minutes [27], drawing concerns on speech-related events to more likely be "superspreader" [27]. Figure 26 shows the aerosols aggregated over a 3-minute window, set to depict a room with adequate ventilation rate. Given a room with a low ventilation rate, we can tweak the parameter win_size to longer duration (e.g., $win\_size = 15$) to simulate aerosol particles lingering in the air much longer.

Key Takeaway: Our aerosol prediction mechanism is a two-step process of predicting aerosol amounts from the detected activity and loudness and aggregating aerosol amounts based on an adjustable parameter that can reasonably represent the dissipation rate in an indoor room. At this point, it is important to emphasize that the ability to predict airborne transmission risks requires knowing both the rate of aerosol generated from human activities and the rate of aerosol dissipation. Our work is intended to work in tandem with indoor air quality systems that can better provide dissipation rates. We discuss the implications of our findings in Section 8.

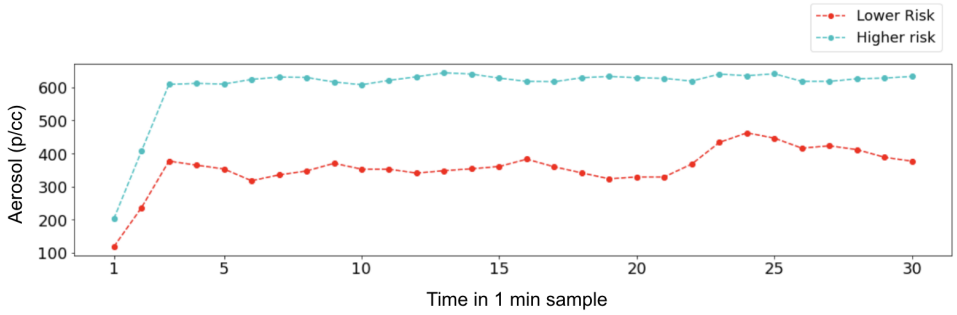


Fig. 26. Accumulated rate of aerosol over a 3 minute window for lower and higher-risk activities.

8 DISCUSSION

We proposed a low-cost mobile-sensing system to estimate aerosol generated from detecting human respiratory activities commonly occurring indoors. Here, we discuss the implications of our findings.

8.1 Impact of Analyzing Speech Content

Inspired by prior work, which found the loudness of activity to increase linearly in aerosol emission [9], our aerosol regression model currently uses dB level (*activity_loudness*) as a feature. However, much research in understanding aerosol sources and chemical compositions has found that vowel content in speech can further affect aerosol emission [9]. Today, advanced speech recognition models can analyze different vowel contents, including proposing novel obfuscation algorithms to detect vowel slices while preserving privacy [60]. By training the aerosol regressor model using all our privacy-preserving features, our preliminary findings shed light on key features that allow us to extend this implementation to more precisely estimate the rate of aerosol generated from human speech using features representative of vowel content.

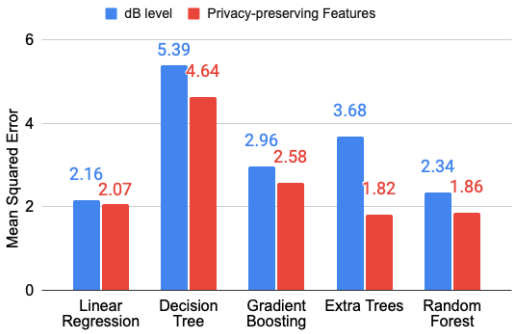


Fig. 27. Using all privacy-preserving features to predict aerosol amounts.

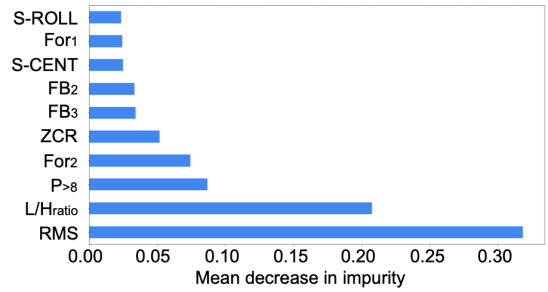


Fig. 28. Features representing speech content as significant predictors of aerosol prediction.

Figure 27 compares the mean squared error using different ML algorithms, with results generally showing a trend of lesser errors with more privacy-preserving features. Further, in Figure 28, a handful of the strongest features are correlated with vowel content. Specifically, a speech will consist of voiced and unvoiced sounds. We conjecture that L/H_{ratio}, ZCR, $P_{>8}$ are significant predictors as they represent high-frequency components

in speech. We also conjecture that formants and filter banks are significant predictors as voiced/vowel sounds are more periodic than unvoiced sounds [61]. Validating these hypotheses will require extensive cleanroom experiments where participants speak with varying clarity through the duct connected to the Condensation Particle Counter.

8.2 Limitations

While our prototype demonstrates the feasibility of using acoustic sensing to recognize common respiratory activities and estimate their corresponding aerosol generation, there are several limitations to the current system. First, we need to extend the system to support various activities. Currently, our prototype implements three human activity recognition capabilities, chosen as the most common respiratory-typed activities, to demonstrate system evaluation with different levels of aerosol generation. In practice, the system must and can support activities such as laughing, singing, and eating and model their aerosol generation. The broader goal of this work is to model and understand the causality between aerosol and acoustic features. This will enable us to predict aerosol for any human activity using audio signals. As we rely on ambient acoustic sensing, detecting very soft sounds, such as heavy breathing, could be challenging. We can integrate wearable-based sensing with our system to sense such soft sounds. Additionally, we must distinguish between similar activities like singing and speaking in a privacy-preserving manner, which becomes challenging. Thus, we plan to delve deeper into the speech content to accurately predict aerosol emissions while ensuring privacy.

Second, *AeroSense* is designed to work in an office or academic environment, and its performance might be affected when deployed in high levels of ambient noise interference scenarios. This will significantly affect the performance of activity detection models, mask detection, voice liveness detection, and even the localization algorithms. High mobility might also affect active source estimation accuracy, leading to overestimating people performing respiratory activities.

Third, aerosol dissipation is a complex relation to the model as it depends on a wide range of factors like temperature, humidity, ventilation, and surfaces in the room. In this work, we use a window size of 3 minutes to aggregate the prior aerosols, and after that, we do not account for previous aerosols. We need information about many different parameters to accurately model aerosol dissipation, making it extremely challenging. For the scope of this work, we do not delve into the chemistry behind the aerosol dissipation, and only focus on the aerosol generation from human activities. In the future, we can integrate advanced aerosol dissipation models into the proposed system.

9 RELATED WORK

Our paper builds upon a significant body of prior work in human activity recognition and audio sensing. Human activity recognition has a rich body of work with different types of sensors such as using IMU [62], gyroscopes [63], microphones [8, 64], and their combinations [13, 65] to detect activities. These sensors can be leveraged from smartphones [64], wearables [13], and ambient environment [8, 65]. Using specifically audio-based techniques, researchers have successfully developed applications that distinguish everyday human activities down to their breathing patterns. When mask-wearing became mandatory during the COVID-19 pandemic, Adhikary *et al.* employed a microphone to monitor breathing in the user's mask [66]. Mohamed *et al.* developed a CNN-based model to detect mask-wearing using a full audio spectrogram [14]. ApneaApp, developed by Nandakumar *et al.*, uses the same audio signals to detect breathing phases for sleep apnea detection [67]. Laput *et al.* uses audio data for real-time detection of a wide range of activities, including speaking, coughing, laughing, and snoring [64]. Recently, BreathEasy [11] proposed the notion of using audio to detect activities and estimate aerosol emissions. However, unlike us, the work did not prototype a system or evaluate its efficacy using user studies.

Audio sensing, which also has a rich history, suffers from privacy concerns when approaches use full audio spectrograms, which would include analyzing human speech. The challenge of reducing privacy invasion has seen possible speech obfuscation workarounds, such as selective sampling for non-speech periods to detect cough and sneeze or suppressing the full audio spectrogram to pick up on breathing sounds. Larson *et al.* presented a system that allows cough sounds to be reconstructed from the feature sets that prevent speech from being reconstructed intelligibly [42]. [68] uses short period of samples to detect sounds related to respiratory symptoms (cough, sneeze, sniffle, throat-clearing) while protecting a user's privacy by not recording raw acoustic data. These techniques proved applicable in airflow sensing where the authors only used the low-frequency spectrum in audio spectrogram [17] to disregard much of human speech lying in the mid and high-frequency bands (500Hz-2kHz). However, these approaches did not focus specifically on audio-based aerosol sensing.

10 CONCLUSIONS AND FUTURE WORK

This paper presented *AeroSense*, a novel privacy-preserving audio-sensing approach that accurately predicts the rate of aerosol generated from detecting the kinds of human respiratory activities and determining the loudness of these activities. Unlike many existing efforts in improving indoor air quality by sensing common indoor air pollutants, our work focused explicitly on sensing aerosol generated from human respiratory activities, which commonly occur in indoor settings and contribute to airborne virus spread. With privacy-first as our key design choice, our work employed a privacy-preserving pipeline of extracting non-reconstructible features to detect common respiratory activities and determine activity loudness from active human sources. Our experimental user studies showed the efficacy of audio sensing for aerosol estimation in real-world settings.

AeroSense is the first step in estimating the risk of airborne transmission using ambient audio sensing. There are several directions for future work. First, *AeroSense* can be extended to detect a broader set of aerosol-generating human activities. For example, studies have shown that singing generates more aerosols than talking. Extending *AeroSense* ś activity detection to address additional activities like singing is an avenue of future work. Further, *AeroSense* ś transmission risk depends not only on aerosol generation but also on the ventilation in an indoor space. Since more ventilated spaces can reduce airborne transmission risk, combining *AeroSense* with audio-based ventilation monitoring systems such as FlowSense [17] is another avenue for future work. Alternatively, *AeroSense* can be integrated into BMS that provide ventilation monitoring and control. In this case, BMS can dynamically enhance ventilation in spaces deemed higher risk by *AeroSense*. The design of such integrated techniques is left to future work.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their suggestions for improving the paper. This research was supported in part by NSF grants 2211302, 2211888, 2213636, 2105494, and US Army contract W911NF-17-2-0196. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Environmental Protection Agency. Indoor Air Quality - What are the trends in indoor air quality and their effects on human health? https://www.epa.gov/report-environment/indoor-air-quality. Online; accessed 14 November 2021.
[2] World Health Organization. Roadmap to improve and ensure good indoor ventilation in the context of covid-19. 2021.
[3] Centers for Disease Control, Prevention, et al. Ventilation in schools and childcare programs. how to use cdc building recommendations in your setting, 2021.
[4] Reopening of schools and universities. https://www.ashrae.org/technical-resources/reopening-of-schools-and-universities. Online; accessed 23 January 2022.
[5] Kristin L Andrejko, Jake M Pry, Jennifer F Myers, Nozomi Fukui, Jennifer L DeGuzman, John Openshaw, James P Watt, Joseph A Lewnard, Seema Jain, California COVID, et al. Effectiveness of face mask or respirator use in indoor public settings for prevention of

sars-cov-2 infection—california, february–december 2021. *Morbidity and Mortality Weekly Report*, 71(6):212, 2022.
[6] Condensation particle counter 3775. https://tsi.com/discontinued-products/condensation-particle-counter-3775/. Online; accessed 1 May 2023.
[7] Justin Morgenstern. Aerosols, droplets, and airborne spread: Everything you could possibly want to know. *First10EM blog*, 6, 2020.
[8] Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–28, 2020.
[9] Sima Asadi, Anthony S Wexler, Christopher D Cappa, Santiago Barreda, Nicole M Bouvier, and William D Ristenpart. Effect of voicing and articulation manner on aerosol particle emission during human speech. *PloS one*, 15(1):e0227699, 2020.
[10] Sima Asadi, Anthony S Wexler, Christopher D Cappa, Santiago Barreda, Nicole M Bouvier, and William D Ristenpart. Aerosol emission and superemission during human speech increase with voice loudness. *Scientific reports*, 9(1):1–10, 2019.
[11] Bhawana Chhaglani, Camellia Zakaria, Jeremy Gummeson, and Prashant Shenoy. Breatheasy: Exploring the potential of acoustic sensing for healthy indoor environments. In *Proceedings of the 1st International Workshop on Advances in Environmental Sensing Systems for Smart Cities*, pages 25–30, 2023.
[12] Sima Asadi, Christopher D Cappa, Santiago Barreda, Anthony S Wexler, Nicole M Bouvier, and William D Ristenpart. Efficacy of masks and face coverings in controlling outward aerosol particle emission from expiratory activities. *Scientific reports*, 10(1):1–13, 2020.
[13] Khuong An Nguyen and Zhiyuan Luo. Cover your cough: Detection of respiratory events with confidence using a smartwatch. In *Conformal and Probabilistic Prediction and Applications*, pages 114–131. PMLR, 2018.
[14] Mostafa M Mohamed, Mina A Nessiem, Anton Batliner, Christian Bergler, Simone Hantke, Maximilian Schmitt, Alice Baird, Adria Mallol-Ragolta, Vincent Karas, Shahin Amiriparian, et al. Face mask recognition from audio: The masc database and an overview on the mask challenge. *Pattern Recognition*, 122:108361, 2022.
[15] Valentyn Stadnytskyi, Christina E Bax, Adriaan Bax, and Philip Anfinrud. The airborne lifetime of small speech droplets and their potential importance in sars-cov-2 transmission. *Proceedings of the National Academy of Sciences*, 117(22):11875–11877, 2020.
[16] Shirun Ding, Zhen Wei Teo, Man Pun Wan, and Bing Feng Ng. Aerosols from speaking can linger in the air for up to nine hours. *Building and Environment*, 205:108239, 2021.
[17] Bhawana Chhaglani, Camellia Zakaria, Adam Lechowicz, Jeremy Gummeson, and Prashant Shenoy. Flowsense: Monitoring airflow in building ventilation systems using audio sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–26, 2022.
[18] Wei Wang, Jiayu Chen, and Tianzhen Hong. Occupancy prediction through machine learning and data fusion of environmental sensing and wi-fi sensing in buildings. *Automation in Construction*, 94:233–243, 2018.
[19] Irvan B Arief-Ang, Flora D Salim, and Margaret Hamilton. Cd-hoc: indoor human occupancy counting using carbon dioxide sensor data. *arXiv preprint arXiv:1706.05286*, 2017.
[20] Wolfgang Schade, Vladislav Reimer, Martin Seipenbusch, and Ulrike Willer. Experimental investigation of aerosol and co2 dispersion for evaluation of covid-19 infection risk in a concert hall. *International Journal of Environmental Research and Public Health*, 18(6):3037, 2021.
[21] Jennifer L Cadnum, Heba Alhmidi, and Curtis J Donskey. Planes, trains, and automobiles: use of carbon dioxide monitoring to assess ventilation during travel. *Pathogens and Immunity*, 7(1):31, 2022.
[22] SN Rudnick and Donald K Milton. Risk of indoor airborne infection transmission estimated from carbon dioxide concentration. *Indoor air*, 13(3):237–245, 2003.
[23] Philip Wexler, Bruce D Anderson, Shayne C Gad, PJ Bert Hakkinen, Michael Kamrin, Ann De Peyster, Betty Locey, Carey Pope, Harihara M Mehendale, and Lee R Shugart. *Encyclopedia of toxicology*, volume 1. Academic Press, 2005.
[24] Anikó Angyal, Zita Ferenczi, Manousos Manousakas, Enikő Furu, Zoltán Szoboszlai, Zsófia Török, Enikő Papp, Zita Szikszai, and Zsófia Kertész. Source identification of fine and coarse aerosol during smog episodes in debrecen, hungary. *Air Quality, Atmosphere & Health*, 14:1017–1032, 2021.
[25] Zhe Peng, AL Pineda Rojas, Emilio Kropff, William Bahnfleth, Giorgio Buonanno, Stephanie J Dancer, Jarek Kurnitski, Yuguo Li, Marcel GLC Loomans, Linsey C Marr, et al. Practical indicators for risk of airborne transmission in shared indoor environments and their application to covid-19 outbreaks. *Environmental science & technology*, 56(2):1125–1137, 2022.
[26] National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. Scientific Brief: SARS-CoV-2 Transmission. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html. Online; accessed 4 May 2023.
[27] V Stadnytskyi, P Anfinrud, and A Bax. Breathing, speaking, coughing or sneezing: What drives transmission of sars-cov-2? *Journal of Internal Medicine*, 290(5):1010–1027, 2021.
[28] Nicholas Good, Kristen M Fedak, Dan Goble, Amy Keisling, Christian L'Orange, Emily Morton, Rebecca Phillips, Ky Tanner, and John Volckens. Respiratory aerosol emissions from vocalization: Age and sex differences are explained by volume and exhaled co2. *Environmental Science & Technology Letters*, 8(12):1071–1076, 2021.

[29] Justice Archer, Lauren P McCarthy, Henry E Symons, Natalie A Watson, Christopher M Orton, William J Browne, Joshua Harrison, Benjamin Moseley, Keir EJ Philip, James D Calder, et al. Comparing aerosol number and mass exhalation rates from children and adults during breathing, speaking and singing. *Interface Focus*, 12(2):20210078, 2022.
[30] Sheng Zhang and Zhang Lin. Dilution-based evaluation of airborne infection risk-thorough expansion of wells-riley model. *Building and Environment*, 194:107674, 2021.
[31] Martin Z Bazant and John WM Bush. A guideline to limit indoor airborne transmission of covid-19. *Proceedings of the National Academy of Sciences*, 118(17), 2021.
[32] JP Duguid. The numbers and the sites of origin of the droplets expelled during expiratory activities. *Edinburgh medical journal*, 52(11):385, 1945.
[33] Rajiv Dhand and Jie Li. Coughs and sneezes: their role in transmission of respiratory viral infections, including sars-cov-2. *American journal of respiratory and critical care medicine*, 202(5):651–659, 2020.
[34] Dirk Mürbe, Martin Kriegel, Julia Lange, Lukas Schumann, Anne Hartmann, and Mario Fleischer. Aerosol emission of adolescents voices during speaking, singing and shouting. *PLoS One*, 16(2):e0246819, 2021.
[35] Malin Alsved, Alexios Matamis, Ragnar Bohlin, Mattias Richter, P-E Bengtsson, C-J Fraenkel, Patrik Medstrand, and Jakob Löndahl. Exhaled respiratory particles during singing and talking. *Aerosol Science and Technology*, 54(11):1245–1248, 2020.
[36] Tehya Stockman, Shengwei Zhu, Abhishek Kumar, Lingzhe Wang, Sameer Patel, James Weaver, Mark Spede, Donald K Milton, Jean Hertzberg, Darin Toohey, et al. Measurements and simulations of aerosol released while singing and playing wind instruments. *ACS Environmental Au*, 1(1):71–84, 2021.
[37] Prateek Bahl, Charitha de Silva, Shovon Bhattacharjee, Haley Stone, Con Doolan, Abrar Ahmad Chughtai, and C Raina MacIntyre. Droplets and aerosols generated by singing and the risk of coronavirus disease 2019 for choirs. *Clinical Infectious Diseases*, 72(10):e639–e641, 2021.
[38] FM Javed Mehedi Shamrat, Sovon Chakraborty, Md Masum Billah, Md Al Jubair, Md Saidul Islam, and Rumesh Ranjan. Face mask detection using convolutional neural network (cnn) to reduce the spread of covid-19. In *2021 5th international conference on trends in electronics and informatics (ICOEI)*, pages 1231–1237. IEEE, 2021.
[39] Siyoung Lee, Junsoo Kim, Inyeol Yun, Geun Yeol Bae, Daegun Kim, Sangsik Park, Il-Min Yi, Wonkyu Moon, Yoonyoung Chung, and Kilwon Cho. An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition. *Nature communications*, 10(1):1–11, 2019.
[40] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. Earsense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–13, 2020.
[41] Duy Duong Nguyen, Patricia McCabe, Donna Thomas, Alison Purcell, Maree Doble, Daniel Novakovic, Antonia Chacon, and Catherine Madill. Acoustic voice characteristics with and without wearing a facemask. *Scientific reports*, 11(1):1–11, 2021.
[42] Eric C Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N Patel. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 375–384, 2011.
[43] Peter F Assmann and Quentin Summerfield. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 88(2):680–697, 1990.
[44] T Schroth. New hepa/ulpa filters for clean-room technology. *Filtration & separation*, 33(3):245–244, 1996.
[45] The inverse square law $1/r^2$ and the sound intensity. http://www.sengpielaudio.com/calculator-distance.htm. Online; accessed 16 December 2022.
[46] Alan Y Gu, Yanzhe Zhu, Jing Li, and Michael R Hoffmann. Speech-generated aerosol settling times and viral viability can improve covid-19 transmission prediction. *Environmental Science: Atmospheres*, 2(1):34–45, 2022.
[47] Logan Blue, Luis Vargas, and Patrick Traynor. Hello, is it me you're looking for? differentiating between human and electronic speakers for voice interface security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 123–133, 2018.
[48] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.
[49] PM De Oliveira, LCC Mesquita, S Gkantonas, A Giusti, and E Mastorakos. Evolution of spray and aerosol from respiratory releases: theoretical estimates for insight on viral transmission. *Proceedings of the Royal Society A*, 477(2245):20200584, 2021.
[50] Zu Puayen Tan, Lokesh Silwal, Surya P Bhatt, and Vrishank Raghav. Experimental characterization of speech aerosol dispersion dynamics. *Scientific reports*, 11(1):1–12, 2021.
[51] Netatmo Smart Home Weather Station. https://www.netatmo.com/en-gb/weather. Online; accessed 16 December 2022.
[52] Respeaker microphone array 2.0. https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/. Online; accessed 16 December 2022.
[53] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.
[54] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*,

17(3):261–272, 2020.
[55] Respeaker microphone array v2.0. https://github.com/respeaker/usb_4_mic_array. Online; accessed 1 May 2023.
[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
[57] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.
[58] Ebrahim Nemati, Shibo Zhang, Tousif Ahmed, Md Mahbubur Rahman, Jilong Kuang, and Alex Gao. Coughbuddy: Multi-modal cough event detection using earbuds platform. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2021.
[59] Maxim Markitantov, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, Alexey Karpov, et al. Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges. In *INTERSPEECH*, pages 2072–2076, 2020.
[60] Terence E Taylor, Frank Keane, and Yaniv Zigel. A speech obfuscation system to preserve data privacy in 24-hour ambulatory cough monitoring. *IEEE Journal of Selected Topics in Signal Processing*, 16(2):188–196, 2021.
[61] Feng Huang, Tan Lee, W Bastiaan Kleijn, and Ying-Yee Kong. A method of speech periodicity enhancement using transform-domain signal decomposition. *Speech communication*, 67:102–112, 2015.
[62] Bishal Lamichhane, Ebrahim Nemati, Tousif Ahmed, Mahbubur Rahman, Jilong Kuang, and Alex Gao. A template matching based cough detection algorithm using imu data from earbuds. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04. IEEE, 2022.
[63] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2023.
[64] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 213–224, 2018.
[65] M Pahar, IDS Miranda, AH Diacon, and T Niesler. Automatic non-invasive cough detection based on accelerometer and audio signals. corr abs/2109.00103 (2021).
[66] Rishiraj Adhikary, Tanmay Srivastava, Prerna Khanna, Aabhas Asit Senapati, and Nipun Batra. Naqaab: towards health sensing and persuasion via masks. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 5–8, 2020.
[67] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*, pages 45–57, 2015.
[68] Xiao Sun, Zongqing Lu, Wenjie Hu, and Guohong Cao. Symdetector: detecting sound-related respiratory symptoms using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 97–108, 2015.