VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2026.1.03
Vibrations in Physical Systems, 2026, 37(1), 2026103

Contactless Respiratory Rate estimation using acoustic beamforming and spectral techniques

Abiodun Ernest AMORAN, Dariusz BISMOR

Faculty of Automatic Control, Electronics and Computer Science, Department of Measurements and Control Systems, Silesian University of Technology, 44-100 Gliwice, Poland

Corresponding author: Abiodun Ernest AMORAN, email: Abiodun.amoran@polsl.pl

Abstract Respiratory Rate (RR) is an important physiological parameter for assessing respiratory health and detecting early signs of ill-health. Microphone arrays offer a promising solution for capturing respiratory sounds in enclosed environments for non-contact health monitoring. However, monitoring RR in a closed environment presents significant challenges, especially when multiple subjects are in proximity, leading to overlapping acoustic signals and spatial interference that make signal separation challenging. This study proposes an RR estimation method based on Direction-of-Arrival (DoA) estimation and beamforming techniques to address this challenge. Acoustic signals were recorded using a circular microphone array, and preprocessed with a Butterworth bandpass filter (100–3000 Hz) to suppress noise and preserve relevant physiological information. A time-frequency representation was obtained via the Short-Time Fourier Transform (STFT), followed by computation of the spatial covariance matrix to characterize inter-channel dependencies. DoA estimation was estimated using the Steered Response Power with Phase Transform (SRP-PHAT) method, and Minimum Variance Distortionless Response (MVDR) beamforming with diagonal loading was applied to isolate individual sound sources spatially. The beamformed signals were reconstructed in the time domain using inverse STFT. Post-processing combined spectral subtraction, Wiener filtering, harmonic enhancement, and adaptive gain control. Signal quality was monitored using voice/activity detection (VAD) based on median and median absolute deviation (MAD) thresholds in the log-energy domain. RR was extracted using Hilbert transform-based envelope detection after bandpass filtering in the respiratory frequency range. Results from the proposed method showed a mean absolute error (MAE) of 1.62 bpm and a root mean square error (RMSE) of 1.93 bpm.

Keywords: beamforming, direction of arrival, microphone, respiratory rate.

1. Introduction

Accurate, contactless monitoring of physiological signals—particularly respiration—remains a significant challenge due to the low amplitude and broadband nature of breathing sounds, which typically occupy frequencies below 3–4kHz [1]. These signals are often masked by speech or ambient noise, resulting in low signal-to-noise ratios (SNRs), especially in real-world environments such as hospital wards or smart homes where multiple individuals may be present. In such settings, separating individual respiratory signals requires spatial information to distinguish sources. Microphone arrays offer a non-invasive solution by leveraging spatial filtering techniques such as beamforming, which relies on direction-of-arrival (DoA) estimation to isolate signals from different locations. Even compact arrays, such as those consisting of four microphones, can significantly enhance respiratory signal detection by improving the SNR through directional filtering. For instance, prior work [2] demonstrated that a six-microphone system combined with beamforming could enhance inhalation and exhalation sounds while suppressing background noise in multi-person settings.

Breathing is a quasi-periodic physiological signal with fundamental breathing frequencies typically below 1Hz, but its acoustic representation includes harmonics extending to a few kilohertz. Consequently, a modest sampling rate of 8kHz or greater is sufficient to capture these signals. However, in shared spaces, overlapping frequency content among different individuals complicates signal separation, reinforcing the need for spatial resolution. Common microphone array geometries include linear and circular configurations. Linear arrays, though simpler, provide high resolution in a single plane and suffer from front-back ambiguity. Circular or square two-dimensional arrays enable omnidirectional coverage with more uniform beam patterns [3]. The angular resolution of an array depends on its aperture and the number

of sensors, as well as the wavelengths of the interest, for breathing sounds, which contain acoustic frequencies ranging from 100 Hz up to 4kHz, which imposes constraints on microphone spacing in order to avoid spatial aliasing. With only four elements, the beamwidth is relatively wide, making it difficult to resolve sources that are spaced closely, for example, if they are less than 30° apart. Nonetheless, circular geometries facilitate full-angle scanning and are well-suited to monitoring multiple users.

Recent work has explored advanced hardware platforms such as smart speakers that incorporate dense microphone arrays and active sonar techniques for contactless vital sign detection. For example, [4] used smart speakers equipped with 6–7 microphones operating in the ultrasonic range (18–22 kHz) to extract chest-wall motion. Furthermore, [5] extended this approach using frequency-modulated continuous wave (FMCW) sonar, enabling heartbeat detection up to one meter with sub-beat-level accuracy in multi-user scenarios. Similarly, [2] used a 6-microphone edge array and beamforming to improve non-contact breathing detection by approximately 12% compared to single-microphone systems. In parallel, developments in resonant micro-electro-mechanical system (MEMS) microphone arrays have improved sensitivity to physiological sounds. Piezoelectric microphones with narrowband acoustic resonances tailored to lung-sound frequencies were adopted by [6]. These resonant arrays act as front-end filters, enhancing desired signal components and reducing power requirements. In the same vein, [7] reported improved wheezing detection with such arrays compared to broadband MEMS microphones.

Various DoA estimation techniques have been adopted for respiratory signal localization, including Steered Response Power with Phase Transform (SRP-PHAT) [8], Generalized Cross-Correlation (GCC) [9], and subspace methods like Multiple Signal Classification (MUSIC) [10] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [11]. Once the locations are estimated, beamforming methods such as delay-and-sum (DAS) or Minimum Variance Distortionless Response (MVDR) can be applied. MVDR, in particular, minimizes interference while preserving signals from desired directions and benefits from accurate spatial covariance estimation. In blind separation scenarios, methods such as Independent Component Analysis (ICA) and Independent Vector Analysis (IVA) have been employed. These approaches can exploit spatial diversity across microphones to unmix overlapping signals in the frequency domain, though their performance may be limited by the statistical similarity of breathing signals. Multichannel Wiener filtering or the Least Mean Squares algorithm [12] in the time-frequency domain presents another alternative, provided the spatial characteristics of each source can be estimated.

Researchers have worked to extend the sensing range of microphone-based monitoring systems. A single speaker-microphone acoustic system based on carrierforming and Continuous-MUSIC (C-MUSIC) to estimate RR and heart rate was implemented by [13]. The carrierforming technique was used to enhance signal-to-noise ratio (SNR) by ensuring coherent superposition across multiple subcarriers along the target path, while C-MUSIC was used to detect motion. The channel frequency response (CFR) was measured, and an Infinite Impulse Response filter was used to recover the RR, while a peak-based method was used to detect the heart rate. Similarly, [14] increased the distance of acoustic signal capture from 2 m to 6 m for RR estimation. A phase-based active sonar approach was employed by [15], a smartphone's built-in speaker was used to transmit an acoustic signal and the microphone captured the returning echoes; changes in the echo phase were then used to estimate RR. Similarly, [16] developed a correlation-based frequency modulated continuous wave (C-FMCW) system using a speaker and microphone for home environments. This method captured periodic variations in the reflected audio signal, which is a result of frequent body movements during breathing. The method measures a signal that represents the RR. A FMCW-based approach using a smartphone was implemented by [17], applying the Hilbert transform for envelope extraction and achieving a mean absolute error (MAE) below 0.15 BPM in various conditions.

The Hilbert transform has also been applied for RR estimation. A novel method named Hilbert-Gaussian transform (HGT)was proposed by [18], combining Gaussian average filtering decomposition (GAFD) and the Hilbert transform within a time-frequency analysis framework. This method overcomes the mode-mixing problems associated with empirical mode decomposition (EMD) and ensemble empirical mode decomposition (EEMD). RR was estimated by calculating the standard deviation of the instantaneous frequency obtained from the mixed Hilbert spectrum and the intrinsic mode functions (IMFs). In the same vein, [19] used the Hilbert transform (HT) to process ballistocardiogram signals. RR was then estimated through peak detection and differentiation, and the results showed that the method obtained a mean absolute error of 0.7 BPM. Similarly, [20] embedded a microphone in a facemask, applied bandpass filtering and the Hilbert transform, and reported an MAE of 1.7 beats per minute (BPM).

Many prior studies rely on smart speakers that emit ultrasonic signals and contain embedded microphones. While effective, these systems differ from passive microphone arrays designed solely for recording. This study presents an experimental set-up that employs a four-microphone circular array without active sonar. This system passively records ambient breathing sounds and applies real-time spatial

filtering and signal processing to isolate and estimate each individual's respiratory rate (RR). The approach emphasizes minimal hardware, privacy preservation, and real-time applicability in shared environments.

2. Data collection and methodology

2.1. Data acquisition

The acoustic data were collected in a controlled acoustic environment at the Active Noise Reduction Laboratory, Silesian University of Technology. The recording setup used a circular microphone array comprising four microphones spaced 0.6m apart (along the array circumference), providing an omnidirectional coverage as shown in Fig. 1. Recordings were conducted in a well-padded, echo-free room to minimize reverberation and external noise interference. The ambient noise level during data acquisition ranged between 28.8 and 29.6 dB, as measured using a SVANTEK 979 sound and vibration analyser. Signals were captured using a TASCAM 16x08 audio interface that served as input to ProTools software. This set-up ensured simultaneous, time-synchronized recordings from all the microphones. All data were initially sampled at 44.1kHz to preserve the frequency content relevant to respiratory and vocal signals, but were later resampled to 8kHz to improve computational efficiency. A PC-3000 patient monitor was used to measure the RR during the experiments. The monitor's five-lead electrocardiogram (ECG) configuration was placed on the subjects' chests according to standard placement guidelines. This served as the ground-truth reference for the research.

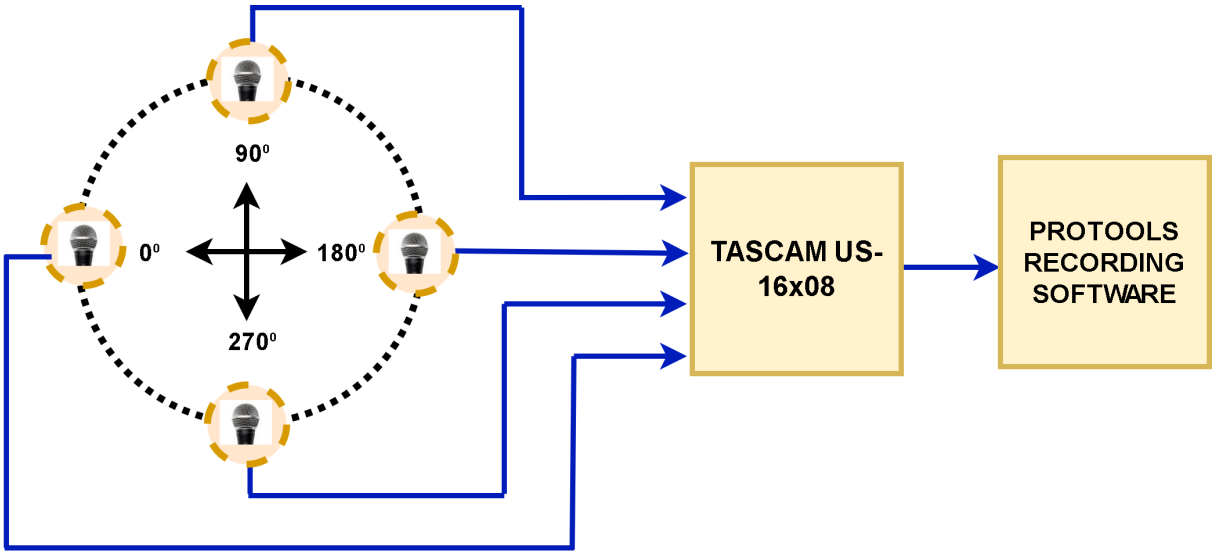

Figure 1. The experimental setup for the proposed methodology.

2.2. Array geometry and signal modelling

The recording set-up employed a four-element uniform circular array (UCA) with a radius of 0.4 m, resulting in a centre-to-centre inter-microphone spacing of 0.6m along the array circumference. The microphones were positioned at azimuthal angles of $0°, 90°, 180°, 270°$, forming a uniform circular array (UCA) geometry. This configuration was chosen to capture low-frequency physiological signals, which are approximately 3kHz, while taking into consideration that higher-frequency components may experience spatial aliasing due to the large microphone spacing. The data were collected from 4 subjects per experiment, with a recording duration of 60 seconds per subject. Moreso, this array configuration provides spatial diversity and uniform angular coverage, which is particularly advantageous for direction-of-arrival (DoA) estimation and beamforming. Assuming a far-field source at an azimuth angle θ, the impinging acoustic wave can be modelled as a plane wave. The corresponding propagation vector is expressed as:

$$\mathbf{u}(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \tag{1}$$

where $\mathbf{u}(\theta) \in \mathbb{R}^2$ denotes the unit direction vector of the source. Let $\mathbf{P}_m \in \mathbb{R}^2$ denote the spatial coordinate of the m-th microphone in the array, and c = 340 the assumed speed of sound. The time delay experienced by the m-th microphone relative to the array center is given by:

$$\tau_m(\theta) = \frac{\mathbf{P}_m^{\mathrm{T}} \cdot \mathbf{u}(\theta)}{\mathrm{c}}. \tag{2}$$

This delay parameter $\tau_m(\theta)$ captures the phase shift introduced by the source direction and serves as the basis for constructing steering vectors in beamforming and spatial covariance estimation.

2.3. Signal pre-processing

The multichannel recordings were first filtered using a Butterworth bandpass filter with a passband of 100--3000 Hz. This frequency range preserves the dominant spectral components of vocal and physiological sounds while attenuating both low-frequency motion artifacts and high-frequency sensor noise. To eliminate phase distortion, zero-phase filtering was applied via forward-backward filtering. Therefore, a time-frequency representation was computed using the Short-Time Fourier Transform (STFT). A Hann analysis window of length L, and hop size H, was employed, yielding complex-valued spectra for each channel:

$$\boldsymbol{X}_m[k,t] = \sum_{n=0}^{L-1} x_n[w] w[n - t\mathrm{H}] e^{-j2\pi kn/N_{FFT}},\ k = 0, \ldots . N_{FFT} - 1, \tag{3}$$

where X_m is the filtered microphone signal at channel m. Only frequency bins corresponding to f_k in [100, 3000] Hz were retained for further processing.

2.4. Spatial covariance and coherence shaping

Let $X[k,t] = X_1[k,t] \ldots \ldots X_m[k,t]^T \in C^m$ denote the STFT vector across m=4 microphones. The spatial covariance matrix at frequency bin f_k was estimated as:

$$\boldsymbol{R}[k] = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}[k,t]\boldsymbol{X}[k,t]^{\mathrm{H}}, \tag{4}$$

where H denotes the Hermitian operator, T is the number of time frames, to reduce spurious correlations introduced by reverberation and microphone spacing, a coherence shaping factor was applied to the off-diagonal elements:

$$Y_{ij}[k] = \mathrm{sinc}\left(\frac{2f_k d_{ij}}{\mathrm{c}}\right),\ \ R_{ij}[k] \leftarrow Y_{ij}[k]\, R_{ij}[k], \tag{5}$$

where d_{ij} is the inter-microphone distance and c is the speed of sound. This procedure enforces physical consistency in inter-microphone coherence, especially at high frequencies.

2.5. Direction of arrival estimation (SRP–PHAT)

Source localization was performed via Steered Response Power with Phase Transform (SRP–PHAT) [8]. For a candidate azimuth θ, the steering vector is defined as:

$$\boldsymbol{a}_k(\theta) = [e^{-j2\pi f_k \tau_1(\theta)} \ldots \ldots e^{-j2\pi f_k \tau_M(\theta)}]^{\mathrm{T}}, \tag{6}$$

To enhance robustness against amplitude variations, the covariance was PHAT-normalized:

$$\widehat{\boldsymbol{R}}[k] = \boldsymbol{R}[k]\ \emptyset\ (|\boldsymbol{R}[k]| + \epsilon), \tag{7}$$

where $\emptyset$ denotes element-wise division and ϵ prevents division by zero. The SRP score was then computed as:

$$P(\theta) = \sum_k \left|\boldsymbol{a}_k(\theta)^H \widehat{\boldsymbol{R}}[k] a_k(\theta)\right|. \tag{8}$$

Angles corresponding to peaks above 30% of the maximum response were selected as the top K candidate directions of arrival (DoAs).

2.6. Minimum variance distortionless response beamforming with diagonal loading

For each estimated DoA θ_s, beamforming was performed using the Minimum Variance Distortionless Response (MVDR) method [10]. The beamformer weights at frequency bin k were given by:

$$w_k = \frac{\check{R}_k^{-1} a_k}{a_k^H \check{R}_k^{-1} a_k}, \quad \check{R}_k = R_k + \lambda I, \tag{9}$$

where

$$\lambda = 0.05 \frac{tr(R[k])}{M}. \tag{10}$$

provides diagonal loading to ensure numerical stability. The beamformed STFT was then obtained as:

$$\mathbf{Y}_s(k,t) = \mathbf{w}_k^H \mathbf{X}[k,t]. \tag{11}$$

The final time-domain signal $\hat{y}_s(n)$ was reconstructed using the inverse STFT with overlap-add.

2.7. Post-filtering and enhancement

A multi-stage enhancement post-separation was implemented to improve the quality of the separated breathing signal. It comprised spectral subtraction, Wiener filtering, harmonic enhancement, and adaptive gain control. The spectral subtraction was used to remove residual noise from each separated source. Each source was normalized, and an estimate of residual noise was calculated and subtracted from the source. The Wiener filtering was applied on a per-frequency basis. The filter calculated the frequency-dependent gain $G(f)$ based on the estimated PSDs of the desired source and noise (equation 12):

$$G(f) = \frac{\Phi_s(f)}{\Phi_s(f) + \Phi_n(f)}, \tag{12}$$

where $\Phi_s(f)$is the estimated power spectral density (PSD) of the desired source and $\Phi_s(f)$is the PSD of noise estimated from silent frames, and $G(f)$acts as a suppression factor usually between 0 and 1, this preserves the source component while attenuating frequency bins dominated by noise. To further highlight periodic components, harmonic enhancement was applied. The fundamental frequency f_o was estimated via Welch's method (range 0.1 – 0.5 Hz) from the PSD. A comb filter was then applied at fundamental and its harmonic multiples, and the STFT of the signal was multiplied by this filter. This step ensures the harmonic structure is preserved, thus improving accurate respiratory rate estimation. These enhancement methods may lead to amplitude inconsistencies across frequency bands, which may be due to array geometry, microphone sensitivity, or noise. Because of this, an adaptive gain control was applied to preserve the spectral structure of the breath signals.

2.8. Robust Voice Activity Detector(VAD)- based signal to Noise Ratio estimation

To avoid inflated SNR from silent frames, a robust VAD method [21] was employed. For each of the audio signals, Frame energies within 20 ms windows with 10 ms hop were computed. The signal $x(n)$ is a bandpass filtered microphone signal $x[n]$, frame energies were computed using window length of L (corresponding to 20 ms) and a hop size of H(corresponding to 10 ms). The frame energy E[t] at frame t is calculated as:

$$E[t] = \frac{1}{L} \sum_{n=0}^{L-1} x^2[n + tH], \tag{13}$$

The log-energies $\ell[t]$ for each frame is then computed to scale the dynamic range, incorporating a small stabilization constant ε, as expressed in Eq. 14:

$$\ell[t] = 10 log_{10}(E[t] + \varepsilon) \tag{14}$$

Frames are then classified as either action or noise by thresholding their log-energies relative to the median log-energy (denoted as $\bar{\ell}$)and median absolute deviation (MAD). Using the default scaling parameters $\kappa_n = 1.5, \kappa_a = 1.0$, frames are labelled according to the following decision rules expressed in Eqs. 15, 16:

$$\text{Noise } if\ \ell[t] \leq \bar{\ell} - \kappa_n \text{MAD}, \tag{15}$$

$$\text{Active } if\ \ell[t] \geq \bar{\ell} + \kappa_a \text{MAD}. \tag{16}$$

If either the resulting active or noise frame set is too small, a fallback threshold using the 20th and 80th percentile is applied instead. Once this classification is done, the average power of the noise frames (P_{noise}) and active frames (P_{active}) are computed over their respective set, as expressed in $P_{\text{noise}} = \frac{1}{|\text{N}|} \sum_{t \in N} E[t]$, and $P_{\text{active}} = \frac{1}{|\text{A}|} \sum_{t \in A} E[t]$, where A and N is the noise and active frame sets respectively. Then the VAD based SNR is computed by:

$$\mathrm{SNR}_{\mathrm{VAD}} = 10\log_{10}\left(\frac{P_{\mathrm{active}}}{P_{\mathrm{noise}}}\right), \tag{16}$$

where P_{active} *and* P_{noise} are averages over corresponding frame sets.

Algorithm 1

VAD-based SNR

Input: Signal, sample rate, frame L, hope H, thresholds κ_n, κ_a
1: $E[t] \leftarrow \frac{1}{L}\sum_{n=0}^{L-1} x^2[n+tH]$
2: $\ell[t] = 10\log_{10}(E[t]+\varepsilon)$
3: $\bar{\ell} \leftarrow \mathrm{median}(\ell)$; $\mathrm{MAD} \leftarrow \mathrm{median}(|\ell - \bar{\ell}|)$
4: $\mathrm{N} \leftarrow \{t: \ell[t] \le \bar{\ell} - \kappa_n \mathrm{MAD}\}$
5: $\mathrm{A} \leftarrow \{t: \ell[t] \ge \bar{\ell} - \kappa_a \mathrm{MAD}\}$
6: $|\mathrm{N}| < 5$ *or* $|\mathrm{A}| < 5$ then
7: use ≤ 20th and ≥ 80th percentiles as fallback sets
8: end if
9: $P_{\mathrm{noise}} \leftarrow \mathrm{mean}(E[t] | \, t \in \mathrm{N})$; $P_{\mathrm{active}} \leftarrow (E[t] | \, t \in \mathrm{A})$
10: return $10\log_{10}(\frac{P_{\mathrm{active}}}{P_{\mathrm{noise}}})$

2.9. Respiratory rate estimation

The enhanced signal was passed through a fourth-order Butterworth bandpass filter with cutoff frequency of 0.1–0.5 Hz to isolate the respiratory frequency band. The Hilbert transform was then applied to compute the signal envelope. The peaks were identified by using the highest thresholds equal to 0.3 times the standard deviation of the filtered envelope, as respiratory cycles manifest as the highest peaks in the filtered envelope. The peaks and troughs were detected by enforcing a minimum separation that corresponds to the breathing rate of 30 breaths per minute. The distance ($d_{\min}$) between successive peaks, expressed in samples, is given by the following equation:

$$d_{\min} = \frac{f_s}{60/30} \approx 0.5 f_s. \tag{17}$$

The inter-peak intervals were computed as the time intervals between successive respiratory cycles, after the detected peaks have been converted to time instants by dividing them by the sampling frequency as expressed in Eq. 18:

$$\triangle t_i = t_{\mathrm{peak},i+1} - t_{\mathrm{peak},i}. \tag{18}$$

Outliers were rejected by using the interquartile range (IQR) criterion. An interval was retained if it lies within the acceptable bounds defined in Eq. 19. Otherwise, any interval lying outside these bounds are excluded from further analysis..

$$[Q_1 - 1.5 \times \mathrm{IQR}, Q_3 + 1.5 \times \mathrm{IQR}], \tag{19}$$

where Q_1 and Q_3 are the 25th and 75th percentiles of the interval distribution and $\mathrm{IQR} = Q_3 - Q_1$. The RR was then computed from the mean of the valid inter-peak interval ($\overline{\triangle t}$) as shown in Eq. 20:

$$\mathrm{RR}(bpm) = \frac{60}{\overline{\triangle t}}. \tag{20}$$

3. Results and discussion

3.1. Beamforming and DoA estimation performance

As shown in Fig. 2, the SRP–PHAT output exhibits prominent peaks in the spatial spectrum, indicating the presence of signal sources at specific angular locations. Four dominant peaks are clearly identifiable, demonstrating the algorithm's ability to resolve multiple sources simultaneously. These peaks correspond to angles with the highest spatial energy, suggesting strong signal coherence from those directions. The SRP–PHAT curve maintains high resolution and stability across the azimuthal plane, with minimal spurious peaks. This confirms the effectiveness of PHAT weighting in suppressing spatial sidelobes and enhancing robustness against reverberation and noise. The detected DoAs, marked in red, align well with the local

maxima of the spatial spectrum, confirming accurate peak detection and localization. These results validate the suitability of SRP–PHAT for multi-source direction estimation in controlled environments, such as the echo-free room used in this study. The uniform angular sweep also highlights the circular array's capability to capture signals from all directions without blind spots or significant spatial ambiguity.

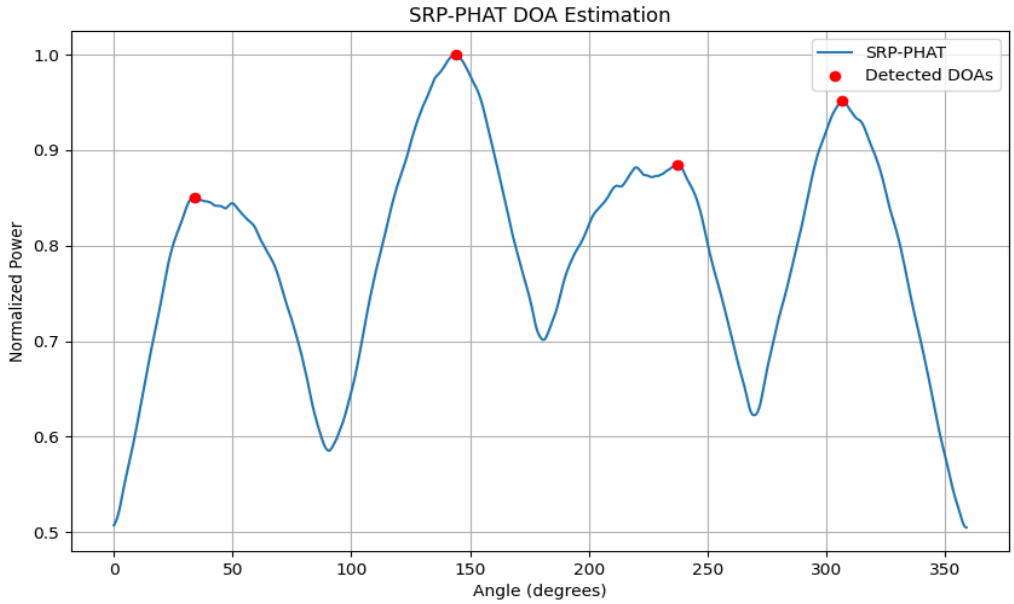

Figure 2. Normalized spatial power spectrum obtained via SRP–PHAT over 0°–360°. Blue curve shows the SRP–PHAT response; red markers indicate detected directions of arrival (DoAs).

In circular array configurations, DoA estimation consistently identified the top four signal directions with sufficient angular resolution for subsequent beamforming.

3.2. Source separation and signal enhancement

Fig. 3 illustrates the progressive stages of the proposed multichannel audio enhancement method for a single microphone signal. The Fig. 3(a) shows the raw audio spectrogram, where low-frequency components below 200 Hz are clearly visible but obscured by broadband noise across the spectrum. The intensity scale indicates that the desired signal is relatively weak compared to background interference. The Fig. 3(b) depicts the beamformed signal obtained via MVDR beamforming. Here, the target source becomes more spatially prominent, with enhanced coherence across the microphone array. While the overall noise floor is moderately reduced, the spectrogram reveals residual interference in mid- and high-frequency regions, consistent with the modest SNR improvement observed in Tab. 1. This reflects the trade-off of MVDR beamforming under imperfect covariance estimation. Finally, the enhanced audio in the Fig. 3(c) demonstrates the combined effect of harmonic enhancement and adaptive multi-band gain. The target spectral components are clearly more pronounced, with a substantial reduction in residual noise and improved contrast relative to earlier stages. Notably, harmonically related patterns are more visible, confirming the efficacy of the harmonic enhancement step. This visual improvement aligns with the quantitative SNR gains reported in the previous section, highlighting the robustness and effectiveness of the proposed pipeline for low-SNR physiological audio signals.

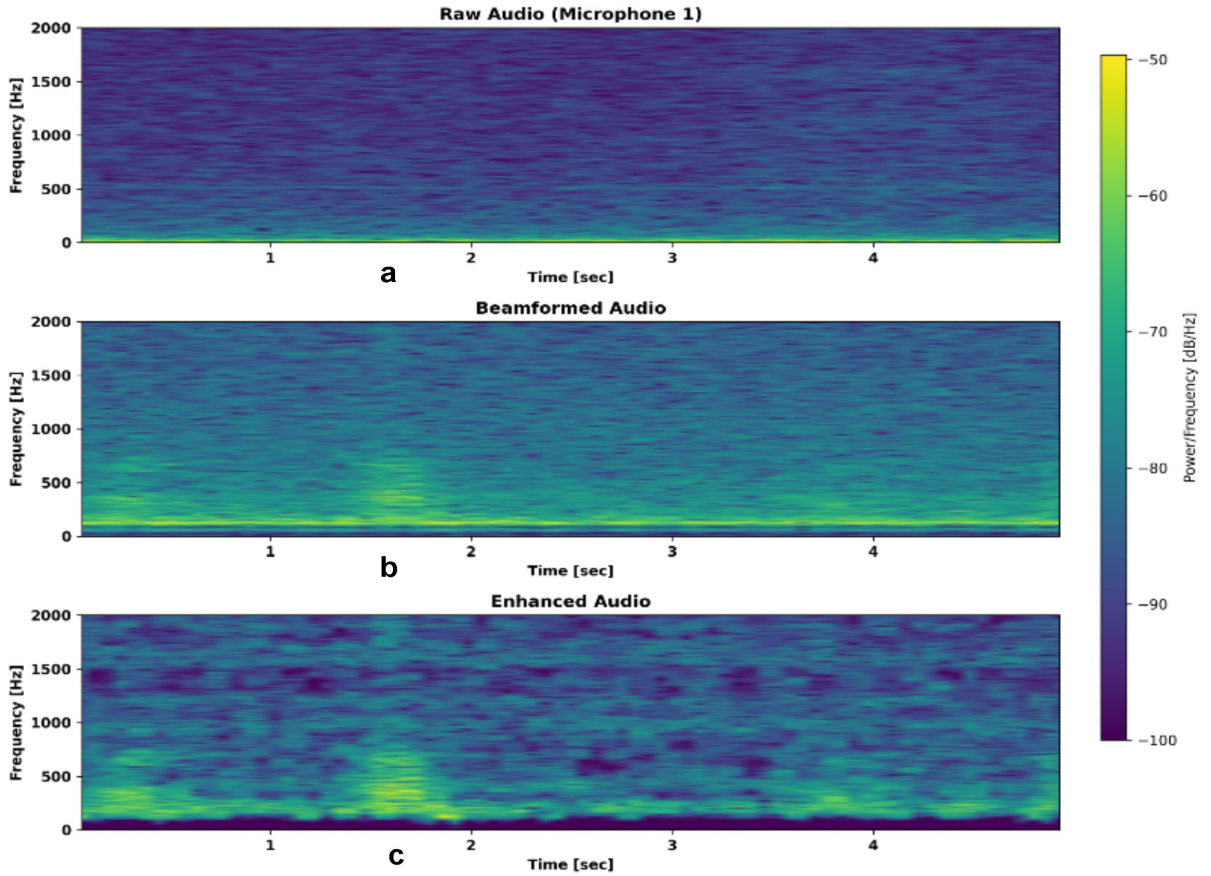

Figure 3. Spectrogram comparison for microphone (a) raw audio: STFT spectrogram of the original multichannel recording, showing frequency content from 0–1000 Hz over a 60 s interval; (b) beamformed audio: spectrogram after applying the array beamforming, highlighting the effect of spatial filtering; (c) denoised audio**:** spectrogram following denoising processing to suppress background noise and interference.

3.3. VAD-based SNR estimation

As described in Subsection 2.8, the signal-to-noise ratio (SNR) was evaluated at multiple processing stages, including beamforming, post-filtering, and respiratory signal extraction. The SNR computation employed a robust VAD-based approach to assess performance, with results illustrated in Fig. 4. The figure highlights two distinct regions: noise segments and active segments, identified by the VAD algorithm using median and median absolute deviation (MAD) thresholds applied in the log-energy domain. Active segments correspond to intervals where the signal energy surpasses the adaptive threshold, capturing the subject's respiratory activity, while noise segments denote low-energy frames that serve as a basis for noise estimation. By explicitly distinguishing these regions, the proposed method mitigates the risk of artificially inflated SNR values due to silent frames, thereby enabling more reliable and accurate performance assessment.

Table 1. SNR Comparison across processing stages.

Mic/Source	Raw SNR(dB)	Beamformed SNR (dB)	Enhanced SNR (dB)
Mic 1/Source 1	6.06	7.74	14.16
Mic 2/ Source 2	7.67	7.70	14.28
Mic 3/ Source 3	5.57	7.70	13.91
Mic 4/ Source 4	7.98	7.72	14.00

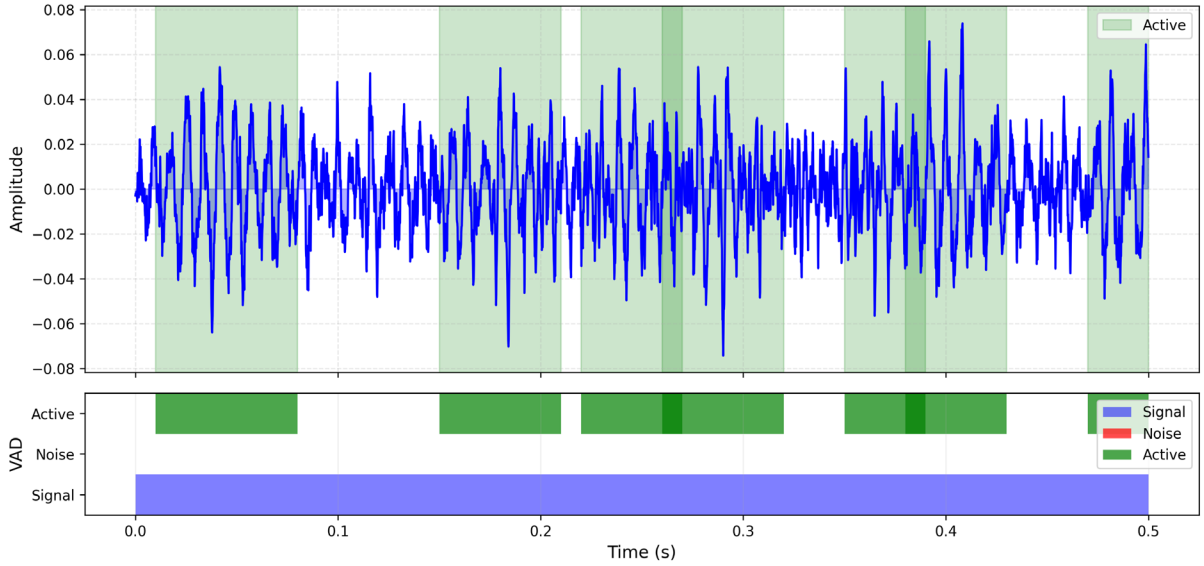

Figure 4. Sample of Voice Activity Detection (VAD) segmentation to calculate SNR.

The SNR values obtained across successive processing stages are summarized in Tab. 1. The raw microphone recordings exhibit SNRs between 5.57 and 7.98 dB, indicating moderate background noise. Following MVDR beamforming, the SNR values ranged from 7.70 to 7.74 dB, showing a slight increase in SNR except for source 4. This indicates that spatial filtering enhances the target subject by suppressing the off-axis noise component. Subsequent post-filtering and harmonic enhancement stages yield a marked improvement in SNR, with values increasing to 14.28 dB. This demonstrates the effectiveness of post-beamforming enhancement in mitigating residual noise while preserving salient periodic components. The improvement is particularly critical for respiratory rate estimation, as the enhanced signals exhibit clearer harmonic structures and reduced contamination from broadband noise. These results indicate that beamforming slightly improves SNR; its combination with spectral and harmonic post-processing achieves substantial gains. This confirms the robustness of the proposed method for source separation in low-SNR environments and underscores its suitability for extracting physiological signals with high fidelity.

3.4 Respiratory rate estimation

Post-denoising, the respiratory signal envelope exhibited distinct amplitude modulations corresponding to breathing cycles as, shown in Fig. 5. Application of the Hilbert transform, followed by peak detection, successfully identified individual peaks and troughs events. The proposed method was able to estimate the respiratory rate (RR) from a separated signal recorded via microphone sensors. In multi-subject scenarios, the approach reliably computed individual RRs with minimal cross-talk, owing to effective spatial separation during beamforming.

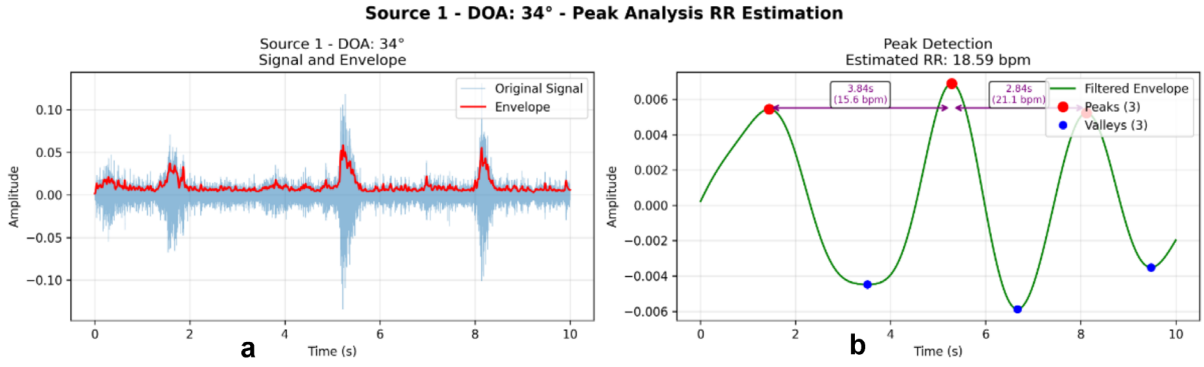

Figure 5. (a) The envelope of the enhanced signal, (b) the peak detection for the RR estimation.

3.6. Performance evaluation

The proposed RR estimation method was evaluated using MAE and RMSE. Experimental results showed that the method achieved an MAE of 1.62 bpm and an RMSE of 1.93 bpm, showing a good performance. As shown in Tab. 1, there is a significant improvement in the SNR (from 5.57 to 14.28 dB) after beamforming and enhancement are applied. This enhancement improved the visibility of the periodic breathing signal significantly, which is important for RR estimation. Furthermore, Tab. 2 shows a comparative analysis between the proposed method and exiting approaches that employed other techniques such as embedded microphones, cameras, smartphones, smart speakers.

Table 2. Performance comparison of respiratory rate estimation methods.

Method Type / Device	Sensor Placement / Modality	Subjects	Environment	Metrics	Values	Sources
Facemask Condenser Microphone	Embedded in facemask	10 healthy adults	Rest, walking, running (3–12 km/h)	MAE, MOD ± LOA, R^2	MAE: 0.5 bpm (rest), 1.7 bpm (overall), up to 3.8 bpm (run); MOD ± LOA: −0.24 ± 5.07 bpm; R^2: 0.94 (rest), 0.9	[20]
Acoustic Wearable (AcuPebble RE100)	Neck sensor	15 healthy adults (prospective), 150 OSA patients (retrospective)	Supine, ambient noise, home sleep study	MAE, RMSD, Bias, r^2	MAE: 1.83 ± 2.09 (prospective), 1.40 ± 1.11 (retrospective); RMSD: 2.78, 2.46 bpm; Bias: 0.63, −0.23 bpm; r^2: 0.87	[22]
Acoustic Respiration Monitor (Radical-7, Masimo)	Adhesive neck sensor	29 healthy adults	Controlled lab, coached/apnea, noise, movement	MAE, Accuracy Error Rate, Absolute Error	MAE: 1.62 ± 0.62 (controlled), 2.19 ± 0.84 (abrupt); Accuracy within ±2 bpm: 76.3% (controlled), 63.0% (abrupt); Absolute Error: 1.67 ± 0.62 bpm	[23]
Smartphone Acoustic (BreathListener)	Near mouth, smartphone mic	30 subjects	Quiet room	MAE, RMSE, Correlation	MAE: 1.6 bpm; RMSE: 2.1 bpm; r: 0.91	[24]
Smart Speaker Sonar (LuckyChirp)	Bedside table (~1m), Google Nest Hub Max	20 subjects	Sleep lab, overnight	MAE, SNR	MAE: 0.48 ± 0.98 bpm; Peak-SNR ~4× FMCW	[25]
Smartphone Sonar (LuckyChirp)	Bedside table (~1m), Pixel 4 / Samsung S6/S7	20 subjects (LuckyChirp), 5 subjects (SonarBeat)	Sleep lab, office, bedroom, cinema	MAE, SNR	Median MAE 0.2 bpm, Mean MAE 0.11–0.33 bpm	[16]
Radar (FMCW, WPCA-MUSIC)	1m from chest, seated	50 healthy adults (25M/25F), 20–25 yrs	EM-shielded room	MAE, SIR	MAE: 1.05 bpm; High SIR	[26]
Radar (FMCW, Adaptive Peak Selection + Kalman)	1.3m height, 1–1.5m from abdomen	41 subjects, 22–72 yrs	Hospital, standing	MAE, SD	MAE: 1.5 bpm; SD: ~1 bpm	[27]
UWB Radar + RGB-D Camera	Room corners, 1m	5 subjects	Lab, non-stationary	Lab, non-stationary	RMSE: 1.03–2.83 bpm (motion-adaptive	[28]
Camera (RGB, Facial Micro-vibration)	1.2m, Raspberry Pi 4/PC	14 subjects	Controlled lighting	MAE, RMSE, Correlation	MAE: 2.017 bpm; RMSE: 2.676 bpm; PCC: 0.930	[29]
Camera (RGB, Optical Flow + PCA)	1.2m, Logitech BRIO 4KHD	14 subjects, 112 recordings	Controlled lighting	MAE, RMSE, Correlation	MAE: 1.244 bpm; RMSE: 1.427 bpm; PCC: 0.968	[30]
Microphone-based beamforming	1m from the subject	4 healthy subjects	Padded Laboratory	MAE, RMSE	RMSE: 1.62 bpm MAE: 1.93 bpm	Our method

4. Limitations and future work

The following are the limitations of this research:

- The study considered only circular array configuration, meaning the results may not be generalized for other array configurations. The size of the microphones could have a significant impact in other configurations.
- The experiments were conducted in an echo-free room, and transferring the system to a real-time environment introduces reverberation and multipath effects, which could make DoA estimation and beamforming challenging.
- The dataset used is relatively small compared with popular databases; however, it is uniquely different in that it incorporates the geometric positioning of multiple microphones rather than using a single microphone.
- The study was implemented as an offline algorithm; a real-time algorithm would require low-latency processing.

Future work will explore deployment on embedded hardware platforms, such as digital signal processors (DSPs) and field-programmable gate arrays (FPGAs), and will investigate deep learning approaches for adaptive direction estimation and signal enhancement. Future research could also develop machine learning models for adaptive DoA estimation in dynamic, multi-user, and reverberant environments, and combine spatial filtering, source separation, and deep feature extraction for robust RR estimation in complex real-time settings.

5. Conclusion

This study presents a robust and modular signal processing method for extracting vital signs— especially respiration rate —from acoustic recordings obtained with a compact circular microphone array. The system integrates spatial processing techniques, including STFT-based covariance matrix analysis, SRP-PHAT for direction-of-arrival (DoA) estimation, and MVDR beamforming, to effectively isolate physiological sources in complex acoustic environments. Post-beamforming enhancement using spectral subtraction, Wiener filtering further improves signal clarity by attenuating residual noise, facilitating the extraction of both low-frequency respiratory components. The Hilbert transform enables envelope analysis, while peak detection allows for reliable RR estimation from low frequency audio signals. The system demonstrates strong potential for remote, non-invasive physiological monitoring applications in healthcare, ambient intelligence, and security domains. Despite the promising results, certain limitations remain. The spatial resolution is inherently limited by the small aperture of the 4-microphone array, particularly at low frequencies. Closely spaced subjects may result in overlapping beams, which could degrade separation performance. Moreover, the system currently operates offline; transitioning to a real-time implementation will require low-latency signal processing and optimized algorithms for DoA estimation and beamforming.

Funding

This work was supported by state budget for science in Poland in 2025 under grant number 02/050/BLM25/0048

Additional information

The authors declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. R.K. Singh, R. Katarya; Recent trends in human breathing detection using radar, WiFi and acoustics; In: Proceedings of the 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), Bhopal, India, 2023; IEEE, 2023, 530–536; DOI: 10.1109/ICRTAC59277.2023.10480776
2. K. Hou, S. Xia, X. Jiang; Buma; Non-intrusive breathing detection using microphone array; In: Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications, Virtual Event, July 2022; ACM, 2022, 1–6; DOI: 10.1145/3539490.3539598
3. A. Szwajcowski, T. Makuch, W. Celniak; An iterative approach to sound source localization based on spherical beamforming; Vibrations in Physical Systems, 2023, 34(2); DOI: 10.21008/j.0860-6897.2023.2.16

4. A. Wang, D. Nguyen, A.R. Sridhar, et al.; Using smart speakers to contactless monitor heart rhythms; Commun. Biol., 2021, 4, 319; DOI: 10.1038/s42003-021-01824-9
5. T. Tran , D. Ma, R. Balan; Remote Multi-Person Heart Rate Monitoring with Smart Speakers: Overcoming Separation Constraint, Sensors, 2024, 24(2), 382;DOI: 10.3390/s24020382
6. A.A. Shkel, L. Baumgartel., E. S. Kim; A resonant piezoelectric microphone array for detection of acoustic signatures in noisy environments; In 2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS), 2015, 917-920
7. X. Ding, D. Clifton, N. Ji, et.al.; Wearable sensing and telehealth technology with potential applications in the coronavirus pandemic; IEEE reviews in biomedical engineering, 2020, 14, 48-70; DOI: 10.1109/RBME.2020.2992838
8. G. García-Barrios, J. Gutiérrez-Arriola, N. Sáenz-Lechón, et. al.; Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps; IEEE Access, 2024, 9, 121549–121560; DOI: 10.1109/ACCESS.2021.3105650
9. J.M. Vera-Díaz, D. Pizarro, J. Macias-Guarasa; Acoustic source localization with deep generalized cross correlations; Signal Process., 2021, 187, 108169; DOI: 10.1016/j.sigpro.2021.108169
10. A. Paulraj, R. Roy, T. Kailath; A subspace rotation approach to signal parameter estimation; Proc. IEEE, 1986, 74(7), 1044–1046; DOI: 10.1109/PROC.1986.13583
11. A. Al-Shoshan; Classification and separation of audio and music signals; In: Multimedia Information Retrieval; IntechOpen, 2020; DOI: 10.5772/intechopen.94940
12. D. Bismor; Leaky partial updates to control a real device casing; Vibrations in Physical Systems, 2022, 33(3); DOI: 10.21008/j.0860-6897.2022.3.04.
13. L. Wang, W. Li, K. Sun, F. Zhang, T. Gu, et.al.; LoEar: Push the range limit of acoustic sensing for vital sign monitoring; Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2022, 6(3), 1–24; DOI: 10.1145/3552496
14. D. Li, J. Liu, S.I. Lee, J. Xiong; LaSense: Pushing the limits of fine-grained activity sensing using acoustic signals; Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2022, 6(1), 1–27; DOI: 10.1145/3493015
15. X. Wang, R. Huang, C. Yang, S. Mao; Smartphone sonar-based contact-free respiration rate monitoring; ACM Trans. Comput. Healthc., 2021, 2(2), 1–26; DOI: 10.1145/3456789
16. T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, B. Dorizzi; C-FMCW based contactless respiration detection using acoustic signal; Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2018, 1(4), 1–20; DOI: 10.1145/3287071
17. A. Chara, T. Zhao, X. Wang, S. Mao; Respiratory biofeedback using acoustic sensing with smartphones; Smart Health, 2023, 28, 100387; DOI: 10.1016/j.smhl.2023.100387
18. Y.D. Lin, Y.K. Tan, T. Ku, B. Tian; A frequency estimation scheme based on Gaussian average filtering decomposition and Hilbert transform: With estimation of respiratory rate as an example; Sensors, 2023, 23(8), 3785; DOI: 10.3390/s23083785
19. O. Linschmann, S. Leonhardt, A. Vehkaoja, C. Hoog Antink; Estimation of the respiratory rate from ballistocardiograms using the Hilbert transform; Biomed. Eng. Online, 2022, 21(1), 54; DOI: 10.1186/s12938-022-01014-x
20. C. Romano, A. Nicolò, L. Innocenti, et al.; Respiratory rate estimation during walking and running using breathing sounds recorded with a microphone; Biosensors, 2023, 13(6), Article 637; DOI: 10.3390/bios13060637
21. R. Yao, Z. Zeng, P. Zhu; A priori SNR estimation and noise estimation for speech enhancement; EURASIP Journal on Advances in Signal Processing, 2016, 101(2016); DOI:10.1186/s13634-016-0406-7
22. R.S. Abdulsadig, N. Devani, S. Singh, et al.; Clinical validation of respiratory rate estimation using acoustic signals from a wearable device; J. Clin. Med., 2024, 13(23), 7199; DOI: 10.3390/jcm13237199
23. M.E. Eisenberg, D. Givony, R. Levin; Acoustic respiration rate and pulse oximetry-derived respiration rate: A clinical comparison study; J. Clin. Monit. Comput., 2020, 34(1), 139–146; DOI: 10.1007/s10877-018-0222-4
24. E.P. Doheny, B.P.F. O'Callaghan, V.S. Fahed, et al.; Estimation of respiratory rate and exhale duration using audio signals recorded by smartphone microphones; Biomed. Signal Process. Control, 2023, 80, 104318; DOI: 10.1016/j.bspc.2022.104318
25. Q.S. Xue, D. Shin, A. Pathak, et al.; Luckychirp: Opportunistic respiration sensing using cascaded sonar on commodity devices; In: Proceedings of the IEEE International Conference on Pervasive Computing

and Communications (PerCom), Pisa, Italy, 2022; IEEE, 2022, 164–171;
DOI: 10.1109/PerCom53586.2022.9762355
26. T. Pei, T. Liao, X. Wan, B. Wang, D. Hao; Spatial blind source estimation of respiratory rate and heart rate detection based on frequency-modulated continuous-wave radar; Sensors, 2025, 25(4), 1198; DOI: 10.3390/s25041198
27. T. Helal, F. Aziz, O. Metwally, et al.; Radar-based respiratory rate monitoring in standing position; arXiv preprint, arXiv:2203.05075, 2022
28. H.C. Hsu, W.H. Chen, Y.W. Lin, Y.F. Huang; Respiratory rate sensing for a non-stationary human assisted by motion detection; Sensors, 2025, 25(7), 2267; DOI: 10.3390/s25072267
29. K.X. Liu, C.Y. Chang, H.M. Ma; Vision-based lightweight facial respiration and heart rate measurement technology; In: New Trends in Computer Technologies and Applications; S.Y. Hsieh, L.J. Hung, R. Klasing, C.W. Lee, S.L. Peng, Eds.; Commun. Comput. Inf. Sci., Springer, Singapore, 2022, 1723; DOI: 10.1007/978-981-19-9582-8_28
30. C. Wiede, J. Richter, M. Manuel, G. Hirtz; Remote respiration rate determination in video data – vital parameter extraction based on optical flow and principal component analysis; In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Porto, Portugal, 2017; SciTePress, 2017, 326–333; DOI: 10.5220/0006095003260333