Hearing Heartbeat from Voice: Towards Next Generation Voice-User Interfaces with Cardiac Sensing Functions

Chenhan Xu[1], Tianyu Chen[1], Huining Li[1], Alexander Gherardi[1], Michelle Weng[1], Zhengxiong Li[2], Wenyao Xu[1]
[1]University at Buffalo, the State University of New York, Buffalo, NY, USA
[2]University of Colorado Denver, Denver, CO, USA
{chenhanx,tchen57,huiningl,ajgherar,mweng,wenyaoxu}@buffalo.edu
zhengxiong.li@ucdenver.edu

ABSTRACT

Voice user interfaces (VUIs) have been adopted in many IoT and mobile devices in daily life. VUIs provide a good user experience with lower-cost hardware (i.e., microphone) and higher throughput (compared with keyboard and touchscreen). Currently, identity authentication and receiving commands are the two most common interactions through VUIs, leaving physiological information in the voice unexploited. Recognizing this untapped potential, we propose VocalHR to extend VUIs beyond voice commands to heart activity sensing without additional hardware. VocalHR is built upon the voice-heart modulation effect, which is rooted in the cardiac activities' impacts on the behavior of the vocal organ during voice production. VocalHR captures voice features of cardiac activity in multiple voice organs and proposes a deep learning pipeline to transform features into cardiac activities. As this is the first study exploring voice-based heart activity sensing, we conducted extensive experiments on 43 demographically diverse subjects to verify the intrinsic link between voice and heart activities. On average, VocalHR can achieve less than 11.1% normalized sensing error on the heart event timing. Our further evaluation shows VocalHR is robust to different microphone specifications and varying speech rates.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Interactive systems and tools**; • **Applied computing** → **Life and medical sciences**.

KEYWORDS

Voice-User Interface; Contactless Sensing; Voice Biometrics; Healthcare

ACM Reference Format:
Chenhan Xu[1], Tianyu Chen[1], Huining Li[1], Alexander Gherardi[1], Michelle Weng[1], Zhengxiong Li[2], Wenyao Xu[1]. 2022. Hearing Heartbeat from Voice: Towards Next Generation Voice-User Interfaces with Cardiac Sensing Functions. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3560905.3568508

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/10.1145/3560905.3568508

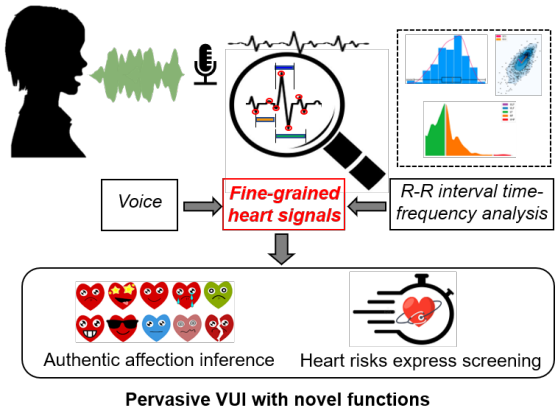

Figure 1: VocalHR enables new functions for VUIs by exporting heart activities from voice.

1 INTRODUCTION

The ever-emerging smart devices in our daily life have witnessed a dramatic growth of voice-user interfaces (VUIs). VUIs enable a more intuitive and low-cost way to operate said devices via natural speech and microphones. They are applied to various scenarios (e.g., message dictation and voice search) to improve operational efficiency. A recent report valued the global VUIs market size to be 13.65 billion dollars in 2020, registering a compound annual growth rate of 21.5% from 2021-2030 [1].

At present, VUIs are tightly bonded to speech commands, while physiological information remains largely unexplored. Considering how promising VUIs is, previous studies recognized this opportunity [2–4] and proposed various personal characteristics in voice for human authentication. Emotion analysis of users' voices is proposed to understand their intent better [5]. Recent studies reveal that our voices also carry biomarkers for diseases (e.g., COVID-19 [6] and Parkinson's diseases [7]). Our voices contain much information that has yet to be scrutinized.

In this paper, we ask the following question: *is it possible to explore cardiac information in voice to endue the VUI applications with new functions?* If we can, a tremendous number of existing and legacy devices with VUIs would be able to perceive human

heart activities without extra hardware. Machines can understand user intentions more thoroughly by utilizing the authentic emotion reflected by heart activities. Voice communication (e.g., phone conferencing and custom services) could be more precise and efficient.

Our work unveils the opportunity of integrating cardiac activity sensing into VUIs by introducing the voice-heart modulation effect. It is based on a known physiological fact that heart activity leads to blood pressure variation, thereby changing the vessel diameter periodically. The vessel deformation happens in the lung and throat. The lungs provide airflow for vocal folds vibration while the throat controls voice production. Considering the correlation between voice and heart rate revealed by previous studies [8], we hypothesize that voice production ought to be influenced by cardiac activity, i.e., the voice carries information about cardiac activities. If our hypothesis holds, devices with VUIs would no longer be limited to explicit commands, and heart activity sensing will be freely available.

Motivated by this vision, we aim to build a system that can transfer microphones into cardiac activity sensors. To achieve our goal, three challenges need to be addressed: (1) Voice is developed for communication. How to discover and extract cardiac activity information from the voice full of semantic information? (2) Each person has a unique vocal system. The cardiac activity information extracted from the voice will be coupled with the user's vocal system characteristics. How to build a model for cardiac activity reconstruction requiring minimum user efforts? (3) How to quantitatively evaluate the reconstructed cardiac activities for various downstream applications?

In this work, we proposed VocalHR, the first systematic framework that utilizes the voice-heart modulation effect to extend VUIs to cardiac activity sensing. We first normalize the voice loudness to compensate for the volume's impact on signal energy. Then, the voice is enhanced by removing the influence of the lip radiation, which is independent of cardiac activity. After that, we analyze the physical model regarding cardiac activity's impact on voice production. Based on the analytical model, we extract representative voice features closely related to cardiac activities' impact on the lung and throat. The analysis also reveals that the model governing voice-heart modulation is coupled with vocal system characteristics. Therefore, we propose a model based on a deep-learning demodulator for cardiac activity reconstruction. Specifically, we utilize a long-short term memory-based filter to extract the cardiac information and then down-convert it to cardiac activities. To configure the voice-heart demodulation model, we design a wavelet decomposition based discriminator for supervision. To evaluate our system, we recruit 43 subjects with results showing 11.02% and 11.08% normalized errors of cardiac event (ventricular depolarization) at low and high heart rate states, respectively.

Our work makes the following contributions:

- We explore a novel voice-based cardiac activity sensing approach. We find that voice carries rich cardiac activity information due to the heart's impact on vocal organs.
- We develop VocalHR, a pervasive cardiac activity sensing system that can be integrated seamlessly with VUIs without additional hardware. Physiological voice features are derived from the voice-heart modulation effect to represent cardiac activity information. A deep learning-based model is proposed to demodulate voice features for cardiac activities.
- We extensively evaluate VocalHR over 43 subjects using multiple cardiac activity metrics. On average, the sensed cardiac activities can achieve an 11.1% normalized timing error of R peaks and a 15.7% normalized error on heart cycle duration.

2 BACKGROUND AND PRELIMINARY

In this section, we introduce the mechanism of phonation and the rationale behind how heart activity can influence voice production. Then, we provide a proof-of-concept study to show the feasibility of VocalHR.

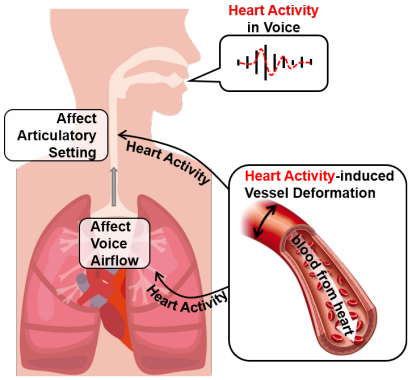

Figure 2: Voice-heart modulation effect is rooted in the heart activity-induced vessel deformation that happens in lung, larynx, and pharynx.

2.1 Phonation and Articulatory settings

The human voice is the result of complex cooperation among multiple articulatory organs [9]. First, the lungs are extruded by the diaphragm and chest cavity to provide the air required by voice production. Then, the air moves through the bronchus and trachea and arrives at the throat, where the vocal folds vibrate and produce basic voice (i.e., air vibration) due to the push of airflow. The air vibration propagates in the airflow and is adjusted and amplified by articulatory organs (i.e., larynx, pharyngeal, and tongue). Finally, the air vibration becomes the voice we hear from our mouths. Humans control the shape, tension, and relative position of these articulatory organs, namely articulatory settings [10], to adjust the airflow and vibration, thereby forming various voices.

2.2 Voice-heart Modulation Effect

The connection between heart and voice production is two-fold. First, the lungs are the organ close to the heart and take a considerable amount of bloodstream from the heart for gas exchange (i.e., Pulmonary circulation) [11]. Second, the articulatory organs in the throat area are surrounded by two carotids, which are the main vessels delivering an enormous amount of blood to the throat and head [12]. As the heart activities-induced blood volume and pressure variation propagate in these vessels, the vessel shape (e.g., diameter) varies accordingly [13].

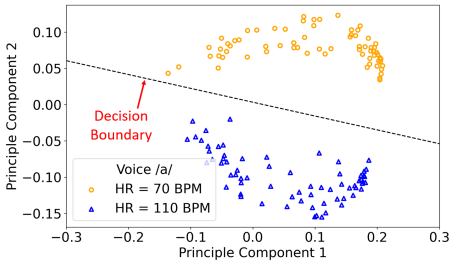

Figure 3: A proof-of-concept of voice-heart modulation effect. A subject will have separable voices for the same phoneme in different cardiac activity states.

Hypothesis: Considering the close link between heart and voice production, we hypothesize that the lung airflow and articulatory settings can be influenced by the vessel shape variation induced by the heart activities, thereby making it possible to reconstruct ECG-like cardiac activity signals from the human voice, namely *Voice-Heart Modulation Effect*. Next, we conduct a proof-of-concept study to support our hypothesis.

2.3 Proof-of-concept Study

To validate the feasibility of the voice-heart modulation effect, we conduct a proof-of-concept study to collect voices corresponding to different heart activity intensities in a room with low ambient noise. As the vowel /ɑː/ is sensitive to the articulatory settings change [14], a subject is told to pronounce /ɑː/ as long as possible in resting state (i.e., low heart activity intensity) during the experiment. In the second trial, subjects do 30 squats to increase the heart activity and repeat the /ɑː/ pronunciation. The voices are collected using a USB microphone with a sample rate of 48000 Hz.

VocalHR Distinction Analysis: If the hypothesis holds, the voices of the subject from different heart activity statuses should be distinguishable. We choose Mel-frequency cepstrum coefficients (MFCC) as well as its two extended variances, BFCC and LFCC, to distinguish voices from different heart activity statuses as they can highlight the voice-range frequency properties and are widely used in voice-based tasks. Figure 3 shows the distinction analysis based on the first two principal components of the normalized features. Each 150-*ms* voice segment yields a data point on the graph. We observe that the voice segments exhibit two clusters, which can be easily separated by a linear decision boundary.

Summary: Our distinction analysis reveals that the voice-heart modulation effect can influence voice production. Therefore, the hypothesis is validated. To further sense cardiac activities using the voice-heart modulation effect, in-depth biological modeling of the voice-heart modulation effect is needed. In the following sections, we will first overview VocalHR's system architecture. Then, we elucidate VocalHR biological voice features that can represent cardiac activities and dive into the data-driven cardiac activity reconstruction.

3 VOCALHR OVERVIEW

In this section, we present the overview of VocalHR. We first illustrate the application scenario, followed by the system architecture.

3.1 Application Scenario

VocalHR extends VUIs to cardiac activity sensing. The system builds on the effect that cardiac activity influences multiple organs involved in voice production. Before the first use, VocalHR requires a one-time user enrollment to profile the user's vocal organ characteristics (as shown in Figure 4). During enrollment, users record their oral reading voices and heart activities simultaneously. The records are used to establish the voice-heart demodulator that will be introduced in Section 5.2. The enrollment setup (illustrated in Figure 8) is intuitive and quick, so it can be done by family practice physicians during a routine visit or provided as a pharmacy service similar to a blood pressure test.

Based on the established model, VocalHR can seamlessly integrate into the existing VUIs to sense cardiac activities when the user is normally interacting with devices such as smartphones and speakers. We further discuss several potential application of VocalHR in healthcare scenarios in Section 10.

3.2 System Architecture

As shown in Figure 4, VocalHR consists of a voice processing module and a cardiac activity reconstruction module. When the user's voice is captured by VUI, its loudness is normalized by pre-processing. The voice-heart modulation effect is then enhanced by removing the irrelevant lip influence of voice. Afterward, the lung, larynx, and pharynx components of voice are extracted, based on which the impacts of cardiac activity on the components are described by the corresponding modulation features. Once the features are obtained, a data-driven demodulator will filter out the cardiac information from the modulation features. The information is finally down-convert to the cardiac activities. During the one-time user enrollment, the user's voices are input to the front of VocalHR, and heart activities are fed to the end. The voice-heart demodulator is configured by this data-driven supervision to profile the user's vocal organ characteristics.

4 VOCALHR PROCESSING SCHEME

In this section, we discuss cardiac activity's influence on voice through an in-depth analysis of the voice production mechanism. The analysis is conducted on the two main components of voice production, i.e., lung airflow-induced vocal folds vibration and articulation. Based on the analysis, we introduce the VocalHR processing scheme, where multiple cardiac activity-related voice modulation features are proposed for cardiac activity sensing.

4.1 Pre-processing

VUIs are utilized by various types of devices, such as smartwatches, home speakers, and robots. Due to human motion and daily activity, users may interact with devices from different distances and directions, which results in voice volume variation. Considering that cardiac activities can influence voice frequency bands and signal power, the impact of voice volume should be eliminated initially. We adopt loudness units relative to full scale (LUFS) as the measure for normalization rather than a-weighted decibel or dB sound pressure level because the normalization with LUFS better correlates with human voice range [15]. We set the normalization to -12 LUFS to keep all voice signals loud enough for further processing

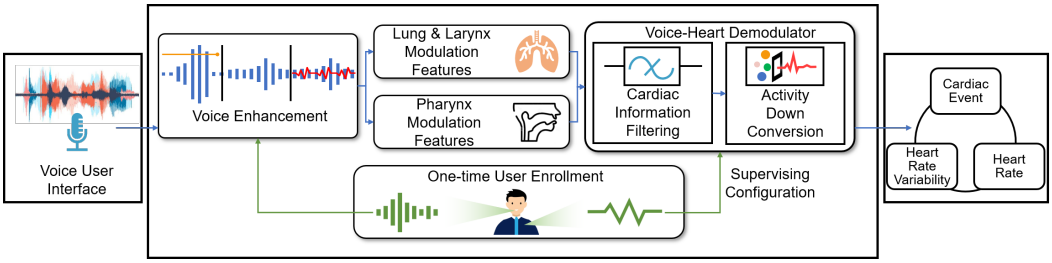

Figure 4: The overview of VocalHR. VocalHR leverages physiological voice features to capture heart's impact on vocal organs and uses these features for cardiac activity reconstruction.

while leaving room for short-term high-volume voice to prevent information loss.

4.2 Enhancing Voice-heart Modulation Effect

The voice production process can be modeled by source-filter model [16], where the voice is considered as the result of filtered vocal fold vibrations. Formally, the voice production could be described as the following transfer function:

$$S(z) = \mathcal{Z}\{s(t)\} = U(z)V(z)R(z), \tag{1}$$

where $S(z)$ is the z-transform of voice signal $s(t)$, t denotes the discrete time point, U, V, R represent the models of lung airflow-induced vibration, articulatory modulation, and lip radiation of the voice wave, respectively. *Since the VocalHR effect happens in the first two parts, it is crucial to enhance these parts and suppress the lip radiation.* In acoustics, lip radiation can be modeled by a piston in a sphere [17], where the sphere models the head and the piston vibration represents the vibration of air between the lips. The model can be formally described as:

$$R(z) = \mathcal{R}_0(1 - z^{-1}), \tag{2}$$

where $\mathcal{R}_0$ is the autocorrelation of $s(t)$ with zero delays. The lip radiation model in Equation (2) has the form of first-order high-pass filter, which can be characterized by +6dB/octave slope (i.e., power increases by 6 decibels when the frequency doubles) [18]. In contrast, the airflow-induced vocal vibration $U(z)$ is considered to be a second-order low-pass filter with −12dB/octave slope. This imbalance makes the higher frequency part of the voice weaker. To balance the frequency spectrum and better capture the VocalHR effect, we enhance the pre-processed voice $s(t)$ by an extra first-order auto-regressive filter:

$$s'(t) = s(t) - \alpha s(t-1), \tag{3}$$

where enhancement coefficient α is set to 0.97 to provide enhancement of +6dB/octave slope. Figure 5 depicts the enhanced voice where the frequency energy is balanced. After the enhancement, the features of VocalHR will be extracted. Next, we will analyze how cardiac activities influence voices and detail the features that could describe cardiac activities' influence.

4.3 Lung-Larynx Demodulation

Lungs provide air that drives vocal folds to vibrate and contain a large number of blood vessels. Thus, the impact of cardiac activities on vessels can be modulated into voice in the lungs. The modulation

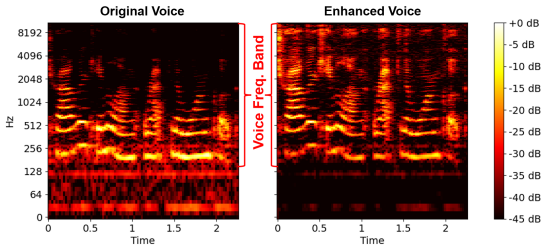

Figure 5: Voice-heart modulation effect is enhanced by compensating the frequency imbalance induced by non-relevant lip radiation.

happens when the lungs are squeezed by the thoracic diaphragm to exhale air for speech, which can be represented as:

$$P_s = \beta\left(\frac{dH}{dt} - \sum_r \frac{dr_l}{dt}\right), \tag{4}$$

where P_s denotes the airflow pressure from lungs (i.e., the pressure under vocal folds), $\frac{dr_l}{dt}$ is the vessel volume (diameter) change due to cardiac activities, H is the volume of lung, β describes the relation between volume change and air pressure that can be derived from Bernoulli's principle. Equation (4) describes that the vessel's volume change due to cardiac activities will influence dynamic lung volume, thereby influencing the air pressure under vocal folds. Knowing P_s, the pressure P that drives vocal folds vibration can be derived on the basis of Bernoulli energy law (air flows through a orifice) [19] as:

$$P = (1 - \frac{a_2}{a_1})(P_s - P_i) + P_i, \tag{5}$$

where P_i is the air pressure caused by the air flow escaping from the vocal folds, $a1$ and $a2$ are the cross-sectional area of glottis entry and exit, respectively. The vocal folds vibration $u(t) = \mathcal{Z}^{-1}\{U(z)\}$ can thus be modeled as a forced mass-spring-damper system [19], where force is from air pressure P, vocal folds is the mass, vocal folds muscles act as spring and damper. This model can be formulated as:

$$M\frac{d^2u(t)}{dt^2} + B\frac{du(t)}{dt} + Ku(t) = P(t), \tag{6}$$

where M, B, and K denote the mass, damping and stiffness of vocal folds, respectively, pressure $P(t)$ is written in a time-dependent form. By substituting Equations (4) and (5) into (6), we obtain a ordinary differential equation about vocal folds vibration and cardiac activity (vessel volume variation). Solving this equation is non-trivial because the parameters (e.g., M, B, and K) are typically

user-dependent and require an intrusive setup to measure. Instead, we utilize vocal folds vibration and its derivatives (i.e., $u(t)$, $\frac{du(t)}{dt}$) as modulation features for further processing. Instead, we utilize vocal folds vibration and its derivatives (i.e., $u(t)$, $\frac{du(t)}{dt}$) as modulation features for further processing (the contribution of these features on performance are evaluated in Sec. 8.1). Since the enhanced voice signal $s'(t)$ contains resonance components of $u(t)$ due to articulatory modulation $V(z)$ (see Equation (1)), the main steps to extract $u(t)$ from $s'(t)$ are:

- Find t_v where the effect of $V(z)$ is strong on $s'(t)$
- Use linear predictive coding (LPC) to model articulatory modulation $V(z)$ around t_v that is found in the first step
- Apply inverse filtering on $s'(t)$ based on the LPC obtained in the step 2

The above steps for $u(t)$ extraction are well-established in acoustics. We adopt the PSIAIF algorithm [20] that can provide precise $u(t)$ extraction by iteratively applying these steps.

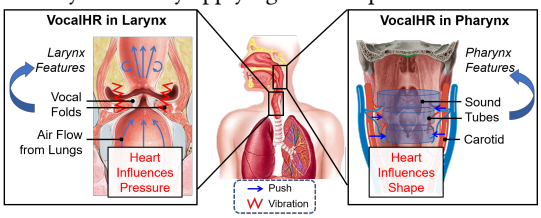

Figure 6: VocalHR's voice features and the underlying biological mechanisms.

4.4 Pharynx Demodulation

In addition to vocal folds vibration, the modulation $V(z)$ in the voice production model is also influenced by cardiac activities. The rationale behind it, as mentioned in Section 2, is the articulatory organs in the throat (mainly the pharynx) could be influenced by the diameter variation of the two carotid arteries. Therefore, based on the acoustic tube model of the pharynx [21], we propose an equation governing the pharyngeal tubes and carotids diameters, which can be denoted by:

$$r_p^n = r_v^n - k2r_c^n, \tag{7}$$

where r_p^n is the effective diameter of the n-th pharyngeal acoustic tube, r_v^n is the ideal diameter of the same tube controlled by muscle according to the $s(t)$, r_c^n is the diameter of the carotid aside the tube, and k is the elastic parameter of the muscle between carotids and pharynx. The constant 2 in the above equation is added because there are two carotid arteries. The above equation describes that heart activity drives the carotid arteries to "extrude" pharynx. As the pharynx is modeled as a chain of several acoustic tubes, reflection exists when sound propagates through the junction of two adjacent tubes. The reflection coefficient k_n can be calculated as:

$$k_n = (r_p^{n+1} - r_p^n)/(r_p^{n+1} + r_p^n). \tag{8}$$

Intuitively, the reflection coefficient is hard to be obtained from the voice $s'(t)$. However, the LPC analysis and voice resonant correlate highly with the reflection coefficient [1], which provides an opportunity to profile the cardiac activity's impact on the voice. Specifically, VocalHR utilizes LPC Coefficients as the pharyngeal features of VocalHR. We use a 30-tube model (i.e., a 30-order LPC, see Section 6.3.1 for details) to have a fine-grained profile of VocalHR (see Sec. 8.1 for the impact of these features to system performance).

[1]See voice analysis literature [22] for more details.

5 CARDIAC RECONSTRUCTION VIA VOICE-HEART DEMODULATION

To demodulate cardiac activity from the extracted features, a solvable model of voice features-activity demodulation need to be established. The analysis in Section 4 reveals the non-triviality of solving the modulation as user-dependent physiological parameters are involved. The recent advances in machine learning provide us an opportunity to solve the problem, i.e., *building personalized demodulator using data-driven infrastructures*. In the rest of this section, we will first analyze the challenges of realizing such a data-driven model, followed by the elucidation of our proposed voice-heart demodulator.

5.1 Design Considerations

Section 4 formally models cardiac activities' influence on voice production. We represent this modulation using an abstract function Λ, i.e., $I = \Lambda(C)$, where I denotes the voice features, C is the cardiac activity. To obtain a voice-heart demodulation model $\lambda_w(I) \approx \Lambda^{-1}(I)$ with learnable parameters w determined through a data-driven approach, we have the following two main considerations.

Consideration 1: Demodulator Architecture. To approximate Λ^{-1}, our demodulator λ_w should have a good learning capability. Studies have revealed that a well-designed structure can benefit learning capacity (e.g., convolutional neural network on image-related tasks) more than simply adding extra neural layers [23, 24]. As the demodulator takes non-IID time-vary cardiac features and outputs heart activities, we consider structures that can utilize historical information for better time-series data processing. In addition, the demodulator should be on a reasonable scale to save the limited computational and storage resources on IoT devices.

Consideration 2: Demodulator Configuration. During user enrollment, the configuration of the demodulator is conducted by minimizing an approximation difference $\mathcal{L}(\lambda_w(I), \Lambda^{-1}(I))$ in terms of the parameters w using user-enrollment voice and heart activity records. Mean squared error (L2) and least absolute errors (L1) are the common choices of the $\mathcal{L}$. However, studies have pointed out that the model configured via L1 and L2 tends to generate blurry results and loses high-frequency details [25] (e.g., R wave, the sharp peak caused by ventricular depolarization) because the error is averaged over multiple enrollment records. Our $\mathcal{L}$ should be carefully designed to optimize the approximation results.

5.2 Voice-Heart Demodulation

An intuitive signal processing method to extract signal-of-interest is *filtering*. Specifically, learnable FIR filters based on convolutional neural networks are utilized in recent studies on cardiac signal extraction [26] due to the absence of hand-crafted filters. However, a large number of stacked convolutional layers are needed to capture the temporal information, which raises the concern of consuming too many resources on IoT devices. In addition, the temporal convolution is designed to process a single time series, which still has

difficulties in processing the temporal VocalHR features vectors. To address these issues, the voice-heart demodulator builds upon the long-short term memory (LSTM), which is considered to be a learnable non-linear IIR filter. The temporal information is captured through recurrent LSTM gates with no limit on the data dimension. Next, We detail the voice-heart demodulator design.

Two-Stage Demodulation. Logically, there are two steps to demodulate cardiac activities from the VocalHR features, i.e., filtering cardiac information from the VocalHR features and down-convert it to the cardiac activity signal (shown in Figure 7). Following the logic, we adopt the encoder-decoder architecture that has been proven to have superior learning capability in many reconstruction tasks [27, 28] as the basic skeleton of the voice-heart transformer. As discussed above, both the encoder and decoder are based on LSTM. In particular, the information filtering conducted by LSTM-based encoder E has the following form:

$$h_t^e, q_t^e = E(\cdots E(E(h_0^e, q_0^e, I_0), I_1), \cdots I_{t-1}), \tag{9}$$

where $h_t^e, t \in [0, T+1]$ is the filtering results (i.e., hidden states of LSTM) that contain cardiac information at time step t, q_t^e represents the encoder LSTM state, T is the total time length of the input VocalHR features. The above equation of the encoder indicates the cardiac information and the LSTM state are passed to the next time step, thereby the historical information is utilized for better filtering. Similarly, the decoder D conducts cardiac activity down-convert as:

$$C_t, h_t^d, q_t^d = D(\cdots D(D(C_0, h_0^d, q_0^d, \mathcal{H}_0), \mathcal{H}_1), \cdots \mathcal{H}_{t-1}), \tag{10}$$

where C_t is the cardiac activity, h_t^d and q_t^d are the decoder LSTM hidden and cell states, and $\mathcal{H}_t$ is the adaptive-weight cardiac information derived from $[h_1^e, \cdots, h_{T+1}^e]$. Different from the encoder, the decoder uses $(C_t \oplus \mathcal{H}_t)$ as the LSTM input and has an extra linear transformation $C_t = \zeta(h_t^d)$ after the LSTM. Next, we will introduce the adaptive weighting in detail.

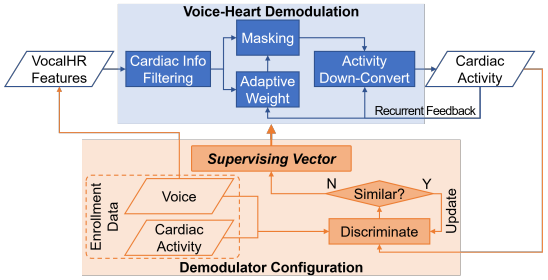

Figure 7: Architecture of the voice-heart demodulation.

Adaptive Weighting of Filtering. Typically, the cardiac information is considered to be $\mathcal{H}_t = \overline{h^e}$ or $\mathcal{H}_t = h_T^e$. However, these considerations assume either the historical cardiac information is of the same importance or the latest information has highest weight, which are strong assumptions. As we depict in Figure 7, VocalHR makes the masking weights of historical information a part of the model parameters w (i.e., attention mechanism [29]). Thereby, the adaptive weighting $\mathcal{H}_t$ can be formulated as:

$$\mathcal{H}_t = \vec{\alpha}_t \times [h_1^e, \cdots, h_{T+1}^e]^\top, \tag{11}$$

where $\vec{\alpha}_t$ is the weight of historical cardiac information. $\vec{\alpha}_t$ is calculated using the current LSTM state h_t^d and all the historical information, which can be formulated as:

$$\begin{aligned} \vec{\alpha}'_t &= [h_t^d \oplus h_1^e, \cdots, h_t^d \oplus h_{T+1}^e] w_{\text{rep}}^\top + b, \\ \vec{\alpha}_t &= Softmax(f(\vec{\alpha}'_t)), \end{aligned} \tag{12}$$

where w_{rep} is the learnable parameter, b is the bias, f is the leaky ReLU function. Knowing the demodulation architecture, we introduce the configuration procedures of the demodulator next.

5.3 Demodulator Configuration

Recently, to avoid the drawbacks of L1 and L2 discussed in Section 5.1, integrating domain knowledge into target functions is widely used for better optimizing the data-driven models. In Cardiology, the fiducial points system (e.g., R wave timing) and derivatives (e.g., Heart Rate Variability) are long and well-studied domain knowledge. In particular, our model λ_w could be considered as a good approximation to Λ^{-1} if the fiducial points and derivatives of the reconstructed cardiac activities are close to their original values. Nevertheless, the calculation of this domain knowledge is not differentiable, making gradient backpropagation and parameter learning impossible. To avoid the dilemma, we utilize a discriminator model as the target function in VocalHR. The discriminator model $\mathcal{D}$ is a binary classification model designed to distinguish the reconstructed and original cardiac activities given the VocalHR features. In contrast, our model will be configured to reconstruct cardiac activities that can not be distinguished (as depicted in the orange area of Figure 7). Now, we can conclude the whole configuration process of the Voice-Heart Transformer as:

$$\arg\min_{\lambda_w} \max_{\mathcal{D}} \mathbb{E}[\log(\mathcal{D}(C, I))] + \mathbb{E}[\log(1 - \mathcal{D}(\lambda_w(I), I))], \tag{13}$$

where our model λ_w and discriminator model $\mathcal{D}$ are configured simultaneously using the enrollment data and only λ_w will be used after the configuration. The configuration is done in a data-driven approach, and the gradient of optimization error (supervising vector) is back-propagated to the demodulator (blue area in Figure 7) to update the learnable parameters. A good discriminator will help our model to demodulate better. Therefore, instead of simply taking cardiac activities, the discriminator decomposes frequency bands in the cardiac activity via convolution-based wavelet decomposition and uses the wavelet coefficients for discrimination.

6 EVALUATION SETUP

In this section, we detail the evaluation preparation and the performance metrics for evaluating VocalHR.

6.1 Evaluation Setup

6.1.1 Experiment Settings. As shown in Figure 8, the experiment setup consists of two simultaneous parts. First, the subjects are required to orally read the selected materials for voice collection. Second, subjects' cardiac activities are captured at the same time as ground truth.

Voice. We use an ordinary USB microphone sampling at 48 kHz for voice recording. The microphone is placed on a desk in front of the subject. The distance between the microphone and the subject is 65 cm. A monitor is used to show experimental subjects the dedicated

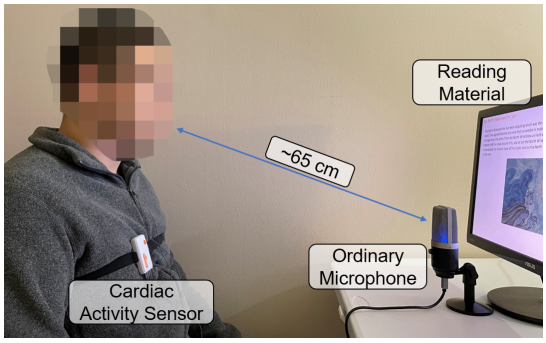

Figure 8: VocalHR experimental setup.

reading materials. We use the standard reading materials (i.e., The Rainbow Passage, Comma Gets Cure, North Wind and Sun, Arthur the Rat, and The grandfather Passage) in voice studies [30, 31] to make sure the collection voices cover most English phonations.

Cardiac Activity. The ground-truth cardiac activities (i.e., Electrocardiogram signal, ECG) are recorded simultaneously as the voice collection using a wearable ECG sensor (Shimmer ECG kit [32]) with a 512-Hz sample rate. Shimmer's integrated clock is synchronized to the voice recording laptop.

6.1.2 VocalHR Collection. 43 subjects (14 females and 29 males) participate in the experiments, including 22 native and 21 non-native English speakers. All the participants are recruited from campus via Emails and advertisements. The experiment is approved by our institutional review board. None of our subjects has voice or heart disease. The average age of the subjects is 21.2 (std=1.7). To get natural voices, first, the subjects are asked to go over the materials and choose four materials for later oral reading. During the experiments, the subjects first orally read two materials to record voices in resting states (i.e., low HR states). Then, their cardiac activity intensity (i.e., heart rate) will be increased through exercises (e.g., push-up, jump, or squats, decided by subjects). As cardio-pulmonary function varies across subjects, subjects are told to stop when they feel short of breath. After a 30-second breath regulation period, the subjects will continue to orally read the remaining two materials. We remark the voices from the remaining two materials as high-HR voices. The mean heart rates of low- and high-HR states are 85.2 (std=10.9) and 100.1 (std=12.1), respectively. A small portion of signals is excluded due to accidental acoustic noises or unstable physical connections. For the oral reading, we ask subjects to use their normal tone and pace in daily life to read the material to get representative results of VocalHR's performance. We use 0.5 seconds as the time length of a sample (overlap=100 ms). The reason for choosing a 0.5-seconds sample length is two-fold. First, a longer sample length tends to include many speech behavior patterns (e.g., utterance habits), which we observe is misleading to the supervision of VocalHR's demodulation. In contrast, a shorter sample is too short for the demodulator to capture temporal information. Based on the sample length, the size of Lung-Larynx features is 4000 (for both $u(t)$ and $\frac{du(t)}{dt}$). The size of demodulated Pharynx features is 208. For each subject, we take a low-HR voice and a high-HR voice for model optimization and use the remaining two oral readings for evaluation. As we collected two readings in resting state and another two readings after exercise, the number of possible train/test splits is four. Therefore, we apply four cross-validations. The average number of samples is 995 (train) and 870 (test) for a subject.

6.2 Performance Metrics

We evaluate the performance of VocalHR from the intra- and inter-cardiac cycle perspectives.

6.2.1 Cardiac Cycle and Event Timing. The heart rate (HR) is yet the most common and intuitive cardiac activity measure. HR sensing has been integrated by many commercial devices (e.g., treadmills and smartwatches) as one of their fundamental features. Typically, the instantaneous HR derived from the duration of a cardiac cycle is reported because of its fast measuring time. Therefore, we use the normalized error of cardiac cycle duration as the fundamental metrics for evaluation, which can be formulated as:

$$Err_{\text{cycle}} = \frac{|T_n^{\text{pred}} - T_n^{\text{ground truth}}|}{T_n^{\text{ground truth}}}, \tag{14}$$

where $T_n^{\text{ground truth}}$ and T_n^{pred} are the duration of n-th ground-truth and reconstructed cardiac cycle, respectively.

As the heart's functionality relies on the periodical and perfect cooperation among multiple heart components, the cardiac cycle and its event timings naturally become the common focus of emerging cardiac activity sensing. Among the cardiac activities, ventricular depolarization (i.e., R peak) is the most significant event that indicates blood pumping [33]. Therefore, R peak extraction is used as the initial step of many cardiac activity analysis methods [34]. We would like to know VocalHR's capability to find R peak timings such that it can facilitate existing analysis methods. Specifically, to exclude the influence from heart rate, we report normalized R wave timing error, which is formulated as:

$$Err_{\text{Rpeak}} = 100\% \times \frac{|t_n^{\text{pred}} - t_n^{\text{ground truth}}|}{T_n^{\text{ground truth}}}, \tag{15}$$

where $T_n^{\text{ground truth}}$ is the aforementioned duration of n-th cardiac cycle in ground truth, t_n^{pred} and $t_n^{\text{ground truth}}$ denote the timings of R wave peaks in VocalHR's prediction and the Shimmer-measured ground truth, respectively.

6.2.2 Heart Rate Variability. In the past decade, studies have revealed that heart cycle variation across multiple cardiac cycles (i.e., heart rate variability, HRV) is a significant indicator of many cardiovascular diseases and mental disorders [35, 36]. The mainstream of HRV analysis can be categorized into time difference-based and frequency-based methods. VocalHR, to extend VUIs to heart activity sensing, is expected to be compatible with existing HRV analysis methods. Therefore, we adopt standard deviation of successive NN interval differences (SDSD) and band power (BP) of HRV as the time and frequency domain metrics. Specifically, the error of SDSD can be calculated as:

$$Err_{\text{SDSD}} = |\text{SDSD}^{\text{pred}} - \text{SDSD}^{\text{ground truth}}|, \tag{16}$$

$$\text{SDSD}^j = \sigma\{T_n^j - T_{n-1}^j | n = 1, 2, \cdots, N\}, \tag{17}$$

where N denotes the total number of cardiac cycles and $j \in$ {pred, ground truth}. The frequency band power are calculated on low-frequency ($f_1 = 0.04, f_2 = 0.15$) and high-frequency ($f_1 = 0.15, f_2 = 0.4$) bands as:

$$Err_{\text{BP}} = 2\left|\int_{f_1}^{f_2} S^{\text{pred}}(f)df - \int_{f_1}^{f_2} S^{\text{ground truth}}(f)df\right|, \quad (18)$$

where $S^{\text{pred}}(\cdot)$ and $S^{\text{ground truth}}(\cdot)$ are the HRV power spectral density functions of the reconstructed and Shimmer measured cardiac activity, respectively.

6.3 System Calibration

We determine the optimal value of key parameters in VocalHR through the following calibration.

6.3.1 Pharynx Model Order. In Section 4.4, we formally described that the pharyngeal vocal settings are influenced by cardiac activities, which can be modeled as sound tube chain variation and captured by Linear Prediction Coefficients (LPC). An appropriate number of sound tube sections (i.e., the order of LPC) is required to capture pharynx settings with minimum information loss. Therefore, we run the cardiac activity reconstruction when the LPC order is set as [10, 20, 30, 40, 50] and record the LPC order where the minimum Err_{Rpeaks} is reached for each subject. During the comparison, only the length of the LPC vector and the input size of the voice-heart demodulator is changed. For each subject, we follow the same setup described in Sec. 6.1.2 without cross-validation. The results reported in Figure 9 show that most of the subjects have their best model when the number of pharynx features sections is 20 or 30. We further observe that the Err_{Rpeaks} reported from 20-section and 30-section models are almost identical (≈12%). Therefore, in the rest of our calibration and evaluation, we set the LPC order as 30.

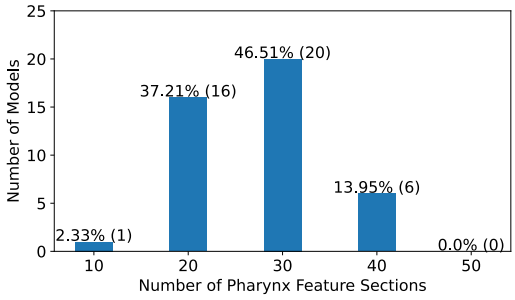

Figure 9: The model distribution by ideal number of pharynx features sections.

6.3.2 Filtering Stack Depth. The VocalHR encoder E conducts a non-linear filtering of the VocalHR modulation features for the cardiac information. The number of stacked LSTM in the encoder is related to the non-linearity of the filtering, thereby being one of the key parameters in VocalHR. Typically, overmuch stacked LSTM increases model configuration difficulty, whereas insufficient LSTM cannot extract meaningful cardiac activity information. Therefore, we test different numbers of stacked LSTMs with other parameters being fixed to determine the extraction stack depth that minimizes Err_{Rpeaks}. The data is prepared following the same setup described in Sec. 6.1.2 without cross-validation. The results shown in Figure 10 indicate that the majority of the configured models have their best performance on four stacked LSTM for filtering. Therefore, we apply four stacked LSTMs in the following calibration and evaluation.

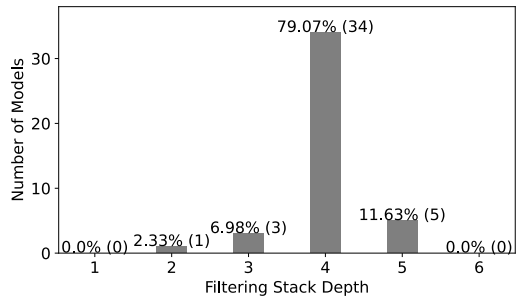

Figure 10: The model distribution by ideal filtering stack depth.

7 OVERALL PERFORMANCE

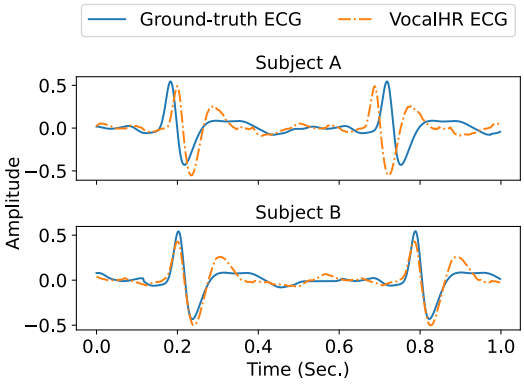

Figure 11: The ground-truth cardiac activity signal and the cardiac activity signal reconstructed by VocalHR.

In this section, we evaluate the overall performance of VocalHR utilizing the metrics detailed in Section 6.2. In Figure 11, we show a sample clip of the cardiac activity signal that VocalHR outputs, compared with the ground-truth signal measured by the sensor. It shows the timing information of peaks is well captured. For further analysis, Figure 12 illustrates the mean normalized R wave timing error Err_{Rpeaks} and cardiac cycle duration error Err_{Cycle} of all subjects for high and low cardiac activity intensities, respectively. We observe that VocalHR has similar performance on both the Err_{Rpeaks} (11.02%, std=4.22% versus 11.08%, std=3.12%) and the Err_{Cycle} (14.63%, std=5.54% versus 16.45%, std=6.17%) for high and low cardiac activity intensities. The low timing errors indicate that VocalHR can perform heart rate sensing well. It is worth noting that the Err_{Rpeaks} is lower than Err_{Cycle} because the R wave offset accumulates when calculating the cycle timing error.

To further understand the quality of the reconstructed cardiac activities, we calculate the cardiac cycle distributions by the two types of error, which are reported in Figure 13. It shows that more than 90% of the VocalHR-sensed heart activities have R wave timing

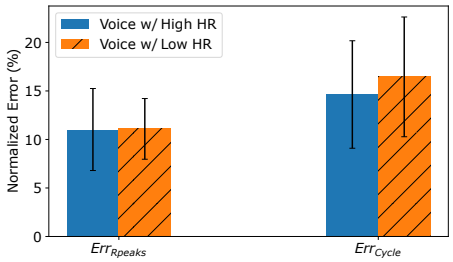

Figure 12: The R wave timing error and cardiac cycle duration error of VocalHR-sensed heart activities.

error less than 17% and cardiac cycle duration error less than 22%. The maximum R wave timing errors of low and high heart rate status are 20.13% and 26.15%, respectively. For the cardiac cycle duration, 32.31% and 33.37% maximum errors are reported for low and high heart rate status, respectively.

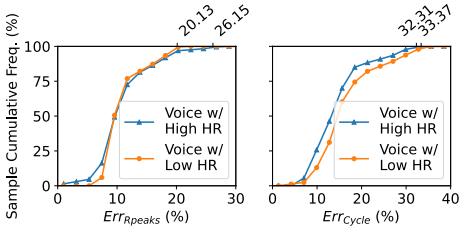

Figure 13: The cumulative frequency of the VocalHR-sensed heart activity by cycle and event timing errors.

Next, we proceed to the heart rate variability (HRV) derivatives. The variability of the successive cardiac cycle (SDSD) is analyzed and reported in Figure 14. The reconstructed cardiac activities report mean *SDSD* as 75.167 ms (std=20.634 ms) and 73.736 ms (std=26.483 ms) for high and low cardiac activity intensities as *SDSD* of sensor-record cardiac activities are 47.197 ms (std=23.122 ms) and 48.159 ms (std=34.603 ms), respectively. The results show VocalHR senses slightly higher successive cardiac cycle variability than the ground truth. We consider it as a normal increase introduced by the cardiac cycle error depicted in Figure 12.

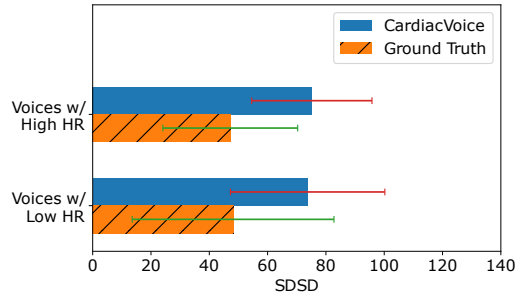

Figure 14: The comparison of standard deviation of successive cardiac cycle differences between VocalHR's reconstruction and ground truth.

In addition, we calculate Err_{SDSD} and report its distribution in Figure 15. The figure shows that VocalHR's Err_{SDSD} are less than 70 for 90% of the subjects without an obvious bias between high and low heart rate status. The mean Err_{SDSD} are 32.399 and 29.801 when subjects are in low and high cardiac activity intensities.

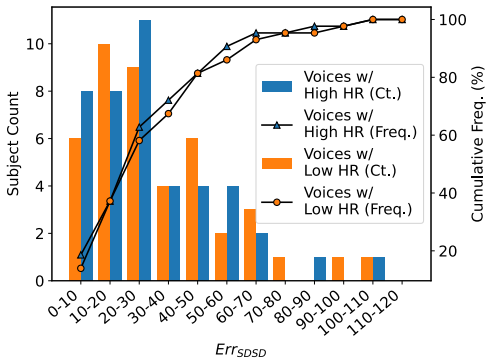

Figure 15: The cumulative distribution of the sensed cardiac activity in terms of Err_{SDSD}

Finally, we compare the reconstructed cardiac activity with the ground truth on HRV band power. As shown in Figure 16, the low-frequency band has mean $Err_{BP} = 1.310 \times 10^{-2} ms^2$ and $Err_{BP} = 1.127 \times 10^{-2} ms^2$ for high and low heart rate status. For high-frequency band, Err_{BP} are $1.344 \times 10^{-2} ms^2$ and $1.269 \times 10^{-2} ms^2$ for high and low heart rate, respectively. The band power of ground truth is $8.88 \times 10^{-2} ms^2$ (low-freq) and $9.89 \times 10^{-2} ms^2$ (high-freq) in low heart rate status. In high heart rate status, the band power of ground truth is $9.21 \times 10^{-2} ms^2$ (low-freq) and $10.85 \times 10^{-2} ms^2$ (high-freq).

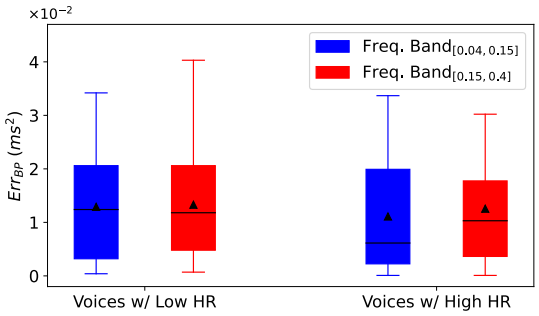

Figure 16: The heart rate variability band power error of VocalHR's reconstruction.

8 CARDINALITY STUDY

To better understand the effectiveness of VocalHR's design, in this section, we systematically remove individual system key components (i.e., independent variables) and gain insights by observing the corresponding system behavior (i.e., dependent variables). Note that our Cardinality study is a type of controlled experiment that includes not only the data-driven voice-heart demodulation part (similar to the ablation study [37] in machine learning system design) but also the biological voice features.

8.1 How Much Does Each Demodulation Feature Contribute?

To verify the effectiveness of VocalHR's biological voice features, we compare VocalHR's performance with different features removed. For a fair comparison, all of these VocalHR variants are retrained using the same parameters and architecture except the pharynx and larynx features. To avoid modifying the data-driven reconstruction, we mask the excluded features with 0 instead of directly removing them. Figure 17 shows the performance degradation caused by masking off each type of feature, compared to the full version VocalHR. We observe removing any type of features will increase the normalized R wave timing error by 9.0% at least. We notice that the performance degradation caused by removing the vocal fold vibration derivatives (16.9%) is greater than that caused by removing vocal vibration (12.2%). This is because the derivatives can better capture the variation and reduce the influence of the fundamental vibration frequency. To conclude, the proposed voice features are closely related to cardiac activity modulation, and each type of feature is non-redundant for cardiac activity reconstruction.

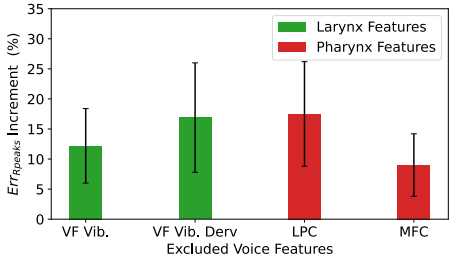

Figure 17: The R wave timing error increment induced by excluding VocalHR features.

8.2 Understanding Cardiac Activity Reconstruction in Voice Data

In this part, we examine VocalHR's design for cardiac activity reconstruction, i.e., the two-stage demodulation, the adaptive weighting, and the demodulator configuration based on the discriminator model. Specifically, we use the default VocalHR setting used in Section 7 as the control group. In addition, we create three experimental groups A, B, and C. In group A, we disable the adaptive weighting and use $\mathcal{H}_t = h_T^e$ instead. In group B, we totally disable the two-step demodulation design, i.e., use simple LSTM for the cardiac activity reconstruction. As for group C, we disable the discriminator and replace it with MSE criteria for the configuration. The absolute changes of Err_{Rpeak} and Err_{Cycle} among groups are reported in Table 1. We observe that the Err_{Rpeak} decreases by 21.3%, 69.7%, and 17.2% for groups A, B, and C, respectively. The Err_{Cycle} also decreases more than 19% for the three experimental groups.

Group	Setting	$Err_{\text{Rpeak}}(\Delta)$	$Err_{\text{Cycle}}(\Delta)$
Control Group	VocalHR Default	11.1% (0)	15.7% (0)
A	w/o Attentional Representation	32.4% (+21.3%)	42.1% (+26.4%)
B	w/o Two-step Reconstruction	79.7% (+68.6%)	86.7% (+71.0%)
C	w/o Discriminator Configuration	28.3% (+17.2%)	35.2% (+19.5%)

Table 1: Absolute changes of Err_{Rpeak} and Err_{Cycle} when individual VocalHR components are disabled.

9 USABILITY AND ROBUSTNESS STUDY

9.1 The Impact of Speech Rate

In the daily use of the voice-user interface, the user may change the rate of speech in different scenarios, which are not expected to degrade VocalHR's performance. Motivated by this usability requirement, we ask three volunteer subjects who have participated in our experiments to do extra oral readings in three different speeds, i.e., slower than their normal speed (slow rate), normal speed (normal rate), and personal fastest speed (high rate). Other experimental settings remain the same. We evaluate the performance using the same model as Section 7, and the results are reported in Figure 18. The figure shows that the HRV band power of VocalHR is not influenced by the speech rate for the three subjects. The observation confirms that the voice features utilized by VocalHR are well-designed to separate the semantic information in voices.

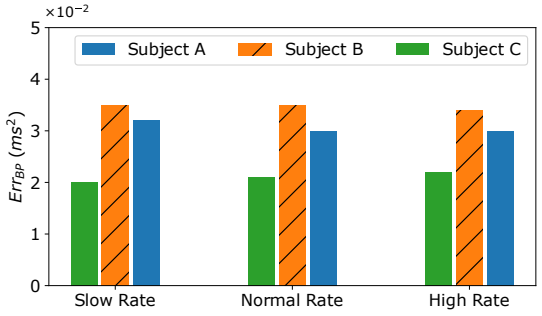

Figure 18: The VocalHR's performance over different user's speech rates.

9.2 The Impact of User Distance & Direction

Considering the pervasiveness of VUI, VocalHR is supposed to have good cardiac sensing capability when people interact with VUIs at various distances and directions. To validate the impact of distance and direction on the performance, we ask the same three volunteers to interact with the microphone in various distance (20cm, 70cm, and 200cm) and directions (azimuth=-65°, -30°, 0°, 30°, 65°; elevation=-60°, 0°, 60°). During the experiment, all other settings are kept the same as that in Section 7. Figure 19 shows the error of averaged HRV band power of the three volunteers. We observe that for each elevation setup, the HRV band power error is reasonably well (mean=3.01) within the normal interaction range ($\leq$ 1 m). When the distance reaches 2 m, the band power error becomes worse (mean=5.02). The direction has little influence on the performance. We think the main reason is the microphone does not pick up the voice from a distance well. The attenuated voices result in an unsatisfactory signal-to-noise ratio for the VocalHR

pre-processing and voice enhancement. It is worth noting that VUIs for distance use usually integrate a directional microphone or microphone array to pick up clear voices. Therefore, we further evaluate VocalHR using a microphone array (UMA-8 [38] with beamforming) that has a similar design as those arrays in smart home assistants. The reported mean error of band power is 3.03 ms^2 at 3 m, which is close to the error within the normal interaction range reported in Figure 19. These results show VocalHR can work well on VUIs deployed in different distances and directions.

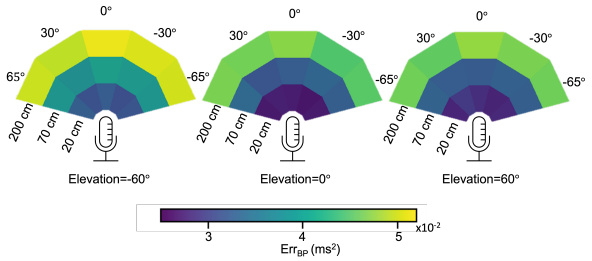

Figure 19: The HRV band-power error when user interacts in various distances and directions.

9.3 The Impact of Voice Sampling Rate

Currently, there are a large number of VUIs running at different sampling rates according to various requirements. The landline phones provide an 8 kHz sample rate for basic voice interaction and communication. The Voice-over LTE (VoLTE) enables a higher sampling rate at 16 kHz. VUIs running on modern smart devices usually have a sample rate of 48 kHz. It is also important to know if VocalHR can accommodate these variances and facilitate pervasive enhanced VUI applications. Therefore, we evaluate the performance of VocalHR under the three most common sampling rates aforementioned. We keep the evaluation setup the same as the other studies in this section except for the sample rate. The results shown in Figure 20 indicate that the Err_{Rpeaks} slightly increases when the sample rate of VUI drops (mean Err_{Rpeaks}= 13.55@8kHz, 12.30@16kHz, 11.61@48kHz). This is because the pharynx demodulation features are slightly influenced by the absence of high-frequency information. Since most of the cardiac information locates below 4 kHz, the performance degradation is limited and VocalHR is robust to different sample rates.

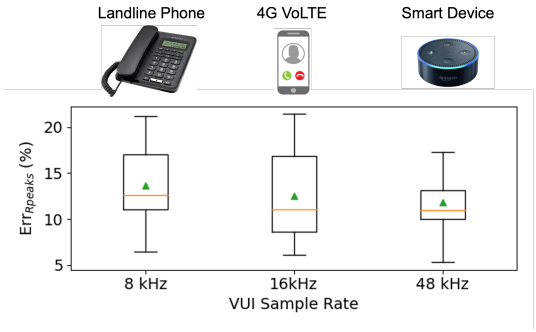

Figure 20: The Err_{Rpeaks} performance under the influence of different sampling rates of VUIs.

9.4 The Impact of Microphone Frequency Response

Besides voice sampling rate, microphones usually have distinct frequency responses, which means different microphones treat frequency components differently. The microphone type and the enclosure are the two main influencing factors of microphone frequency response. To evaluate the impact of microphone frequency response, we tested three microphones carried by different devices, which are a condenser microphone (the same one used in Section 7), an electret microphone (Logitech Pro Gaming microphone), and a MEMS microphone (Poco X3). The frequency responses of the microphones are shown in Figure 21. The mean Err_{Rpeaks} of a condenser microphone, MEMS microphone, and electret microphone are 12.7, 13.2, and 12.9, respectively. We observe that the frequency response has little influence on VocalHR's performance. Therefore, we can conclude that VocalHR is able to support and improve a very large part of the existing VUIs.

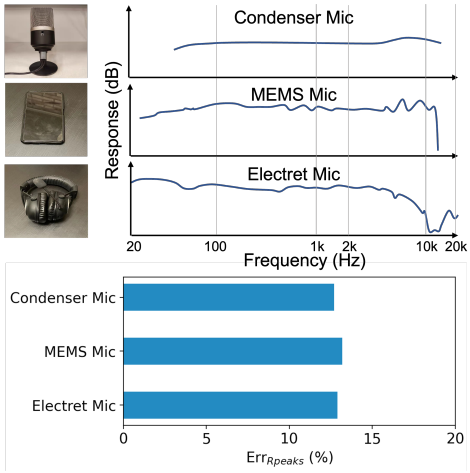

Figure 21: The Err_{Rpeaks} performance under the influence of different microphone frequency responses of VUIs.

9.5 The Impact of Noise

When the user speaks, the generated sound wave propagates in the air media, which is prone to interference from ambient noises. As the noises can make VUI insensitive to minute alterations of the human voice, we are wondering if it can decrease the performance of our system. Therefore, we evaluate our system under three typical types of real-world noises with diverse spectral properties, i.e., home appliances noises, music noises, and talking people noises. Specifically, we place a loudspeaker near the subject and playback the recording of these three types of noises at different loudness levels, i.e., volume varies as 0 (38 dB SPL), 33% (52 dB SPL), 66% (63 dB SPL), and, 100% (76 dB SPL). In the same while, the subject reads the required materials. The collected audio signals are fed into our system. For comparison, we further apply software-based noise reduction technology [39] on the 100% noise group to see how the noise reduction algorithm influences VocalHR. With current experiment setting, Figure 22 shows the results.

Home appliances. The noises from the home appliance (a washer) occupy a narrow frequency band, which has a large overlap with

some voice harmonics. We observe that the Err_{Cycle} increases as the noise gets louder. Because the noise is narrow-band, the performance degradation is limited. We also observe that noise reduction work well on the home appliance noise, and the performance drop is controlled by noise reduction.

Music. The music background noise mainly influences the lower frequency. We see the noise pollute the fundamental frequency band of the subject's voice. Therefore, the performance drop is higher than the simple home alliance noise. Noise reduction can reduce the influence of music noise.

Talking people. The talking noise has a wider frequency band and is totally mixed with the subject voices. In this case, the performance decreases fast when the volume is increased. Different from the music noise, VocalHR can get much less cardiac information from the VUI. In the worst case, noise reduction can still reduce the impact of noisy talking.

To conclude, VocalHR can be robust to common single-tone noises utilizing the derivatives of vocal folds vibration. For noises within voice frequency, the influence can be reduced by noise reduction. For example, voice echo interference can be largely suppressed by adaptive filtering [40]. Furthermore, VocalHR is also compatible with spatial noise cancellation (i.e., isolating the speaking person) if the VUIs are based on microphone arrays.

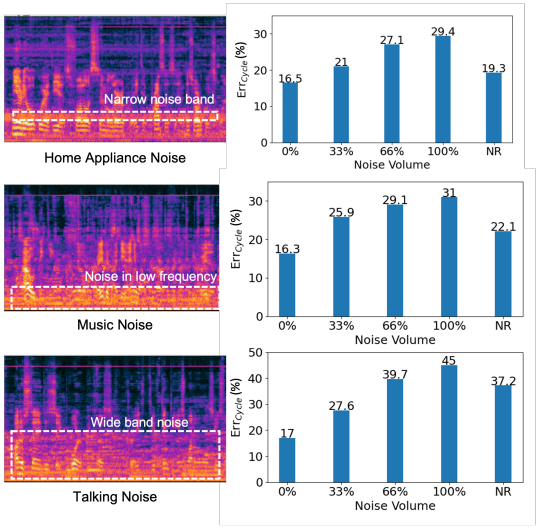

Figure 22: The performance of VocalHR under the influence of different types of noise.

9.6 Impact of Real-world Environment

To evaluate VocalHR's robustness to environmental dynamics, we further test VocalHR in four real-world scenarios, i.e., office, meeting room, warehouse, and shopping mall. We use a hand-held iPhone 12 as the voice collection device to have a typical setup of daily VUI use. The synchronization between the heart activity sensor and iPhone is done by an extra camera recording the start recording event timestamps. During the collection, we ask seven subjects to walk around freely while reading the materials printed on the paper. Other procedures remain the same as stated in Sec. 6.1.2. As depicted in Figure 23, the performance of VocalHR is steady during motion and in various environments (mean=11.27%). VocalHR is robust to the variance of environmental dynamics such as clutter and the presence of echos.

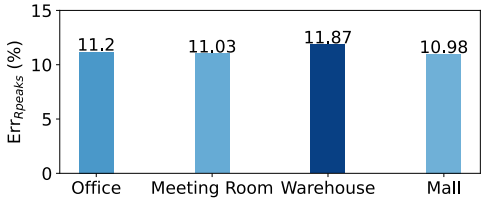

Figure 23: The R wave timing error in various real-world environments.

9.7 Longitudinal Study

Similar to biometrics [41], it is important to prove the permanence of the voice-heart modulation effect to ensure the robustness of VocalHR. To conduct the longitudinal study, we collected data from 10 subjects randomly selected from the original subjects in 100 days. We follow the same data collection protocol as mentioned in Sec. 6.1.2. The R wave timing error and cardiac cycle duration error are shown in Figure 24. The reported Err_{Rpeaks} is 11.27%, std=4.00% versus 10.89%, std=2.95%) for high and low cardiac activity intensities. The Err_{Cycle} is 14.01%, std=6.15% versus 16.00%, std=6.47%. We observe that neither of the errors has a significant increase or drop compared to that reported in Figure 12. These results demonstrate that VocalHR is robust against time change and reveals the permanence of the voice-heart modulation effect.

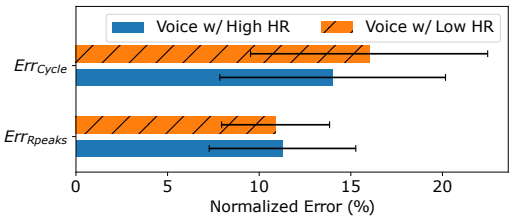

Figure 24: Longitudinal study of VocalHR in 100 days.

10 IMPLICATIONS AND FUTURE WORK

Emergency Healthcare. Cardiac health services are receiving ever-increasing attention as the rising trend of cardiac disease in various age groups. By enabling the voice-based cardiac activity sensing, VocalHR makes an important step toward the telehealth of the heart. The core benefit VocalHR brings is that it requires no extra sensors. Any device carrying a voice-user interface can be used as VocalHR's front-end. The personalized VocalHR model may be integrated into personal health records and shared with the telehealth provider to reduce the difficulty of initial diagnosis. As the research evolves, VocalHR may be used in an emergency (e.g., emergency call) to retrieve the caller's physiological status for quick responses with no extra effort.

Affective Communication. In addition to receiving explicit commands, VocalHR can endue VUIs with advanced affection-related

functions, such as authentic emotion detection and speaking intentions inference. Existing works roughly estimate the outward expression of emotions from speaking content and tones, but users may hide their emotions, and they differ largely in how to express their emotions. Different from these works, VocalHR can be used to detect the user's inner feelings more precisely by monitoring our obtained physiological signals. In this case, VUIs can help interact with machines and people in a more efficient and satisfactory manner. For example, VUIs would control smart home appliances to respond to users' current moods by adjusting the lighting conditions or playing relaxed music automatically. The Advertisers can learn customers' authentic reactions instantly and recommend the most suitable products.

Cardiac Phonetics VocalHR utilizes the heart activity information modulated in the voice to enable heart sensing for VUIs. Considering human voices can be decomposed into smaller unit as phonemes (e.g., vowel and consonant), a potential future work is to understand the entropy of each phoneme in terms of heart information. With this in-depth knowledge, more effective voice processing methods can be developed to enable different weights for phonemes according to the targeted applications and further improve sensing capacity. In addition, future works could investigate the heart sensing capacity of voice through more fine-grained (e.g., under different emotional conditions) and diverse (e.g., from heart disease patients) data.

Fake Voice Detection. There has been a proliferation of publicly available audio manipulation software to disguise one's voice for a variety of objectives from casual fun to safeguarding personal privacy to evading voice biometrics applications and even spoofing other individuals through deep audio fakes. VocalHR has the potential to detect the fake voice by leveraging the extracted cardiac properties. Specifically, we can feed the audio signal to VocalHR and calculate the heart cycle duration and the timing of cardiac events (e.g., R peaks). If these cardiac measurements are not within the normal range, this audio can be regarded as a fake voice disguised by software.

11 RELATED WORK

11.1 VUI-based Biotraits

There is a long history of exploring the bio-information on voice-user Interfaces (VUIs). The past 50 years have witnessed the great success of biometric-based identification on VUIs (e.g., voice authentication [42], and liveness detection [4]). Recently, voice interaction is revealed to carry biotraits associated with a multitude of biological processes. Tao et al. [43] proposed an ensemble learning-based voice emotion recognition system working on real-world audio. Zhou et al. [44] proposed a deep learning-based semi-supervised approach to infer the emotion from voice and validated the proposed approach on a large-scale Internet voice dataset. Inferring speak intention from audio is discussed in a recent work [45]. In addition, smartphone VUIs are explored to detect pathological conditions, such as Parkinson's disease [46] and COVID-19 [47]. Our work shares the same vision as these previous studies and, for the first time, brings cardiac activity sensing onto the pervasive VUIs. It is worth noting that there are several pioneer studies [8, 48, 49] exploring heart rate estimation from voice. They show the feasibility of cardiac activity sensing through VUIs. Our work differs as it senses the intrinsic activities and events in each cardiac cycle utilizing the biological voice features and a data-driven reconstruction model.

11.2 Non-contact Heart Activity Sensing

Current studies of non-contact heart activity monitoring rely on sensing the vibration induced by heartbeats. An idea to get rid of contact sensors on human body is to install vibration sensors on human contact structures to indirectly sense the heart activity (i.e., semi-contact). Studies [50, 51] proposed to embed geophones into bed for heart monitoring during sleep. Bonde et al. [52] enabled heart monitoring for drivers by integrating accelerometer array into car seats. To conduct fully non-contact heart sensing, Liu et al. [53] leveraged Channel State Information (CSI) of WiFi signal for heart rate estimation. Yang et al. [54] proposed to use Received Signal Strength (RSS) as an indicator of heart rate. Zhao et al.[55] developed an RFID-based non-contact heartbeat monitoring system using the WiFi RSS phase. To sense fine-grained heart activity, later studies utilized dedicated waveforms and higher frequencies to increase sensing capability. A 2.4 GHz Doppler radar is used in [56] to capture heart activity fiducial points. Ha et al. [26] proposed to use 77 GHz FMCW radar to capture chest displacement and estimate seismocardiogram signal. Xu et al. [57] proposed a mmWave-based scheme to generate electrocardiogram-like signals via heart electromagnetic field sensing. Besides, [58] utilized light reflection on the face to sense the heart pump period. Differently, VocalHR is a pervasive and low-cost heart sensing system built upon the voice-user interfaces without additional hardware. Based on the heart information modulated in voice, VocalHR is more robust to motion interference and occlusion. Moreover, utilizing the pervasive phone networks, VocalHR can work remotely on the callee side, thereby requiring no internet or smartphone on the sensing end.

12 CONCLUSION

In this paper, we proposed the first voice-based pervasive cardiac activity sensing system, VocalHR, to bridge cardiac health applications to the omnipresent voice-user interfaces (VUIs). We explore heart activity's impact on the voice production process and propose the corresponding biological lung-larynx and pharynx VocalHR features to describe the heart activities in the voices. Then, we propose a novel cardiac activity reconstruction model that can demodulate the cardiac information from features and reconstruct cardiac activities. We propose to use a discriminator with wavelet decomposition to supervise the data-driven demodulator configuration. Extensive experiments show the effectiveness of VocalHR to sense cardiac activities from human voices.

ACKNOWLEDGMENTS

We thank all anonymous reviewers for their insightful comments on this paper. This work was supported by the National Science Foundation under grant No. 2028872 and No. 2050910.

REFERENCES

[1] P. Borasi, H. Jangra, and V. Kumar, "Voice user interface market size, share and analysis: Forecast - 2030." [Online]. Available: https://www.alliedmarketresearch.com/voice-user-interface-market-A12381

[2] J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[3] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David, "Biometric authentication on a mobile device: A study of user effort, error and task disruption," in *Proceedings of the 28th Annual Computer Security Applications Conference*, ser. ACSAC '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 159–168. [Online]. Available: https://doi.org/10.1145/2420950.2420976

[4] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2062–2070.

[5] C. Breazeal, "Emotion and sociable humanoid robots," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[6] A. Hassan, I. Shahin, and M. B. Alsabek, "Covid-19 detection system using recurrent neural networks," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2020, pp. 1–5.

[7] H. Zhang, C. Song, A. Wang, C. Xu, D. Li, and W. Xu, "Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–16.

[8] B. Schuller, F. Friedmann, and F. Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7219–7223.

[9] Z. Zhang, "Mechanics of human voice production and control," *The journal of the acoustical society of america*, vol. 140, no. 4, pp. 2614–2635, 2016.

[10] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.

[11] A. J. Peacock, R. Naeije, and L. J. Rubin, *Pulmonary circulation: diseases and their treatment*. CRC Press, 2016.

[12] D. Holdsworth, C. Norley, R. Frayne, D. Steinman, and B. Rutt, "Characterization of common carotid artery blood-flow waveforms in normal human subjects," *Physiological measurement*, vol. 20, no. 3, p. 219, 1999.

[13] J. M. Meinders and A. P. Hoeks, "Simultaneous assessment of diameter and pressure waveforms in the carotid artery," *Ultrasound in medicine & biology*, vol. 30, no. 2, pp. 147–154, 2004.

[14] D. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3177–3186, 2005.

[15] T. Fischer, M. Caversaccio, and W. Wimmer, "Multichannel acoustic source and image dataset for the cocktail party effect in hearing aid and implant users," *Scientific data*, vol. 7, no. 1, pp. 1–13, 2020.

[16] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[17] U. Laine, "Modelling of lip radiation impedance in z-domain," in *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7. IEEE, 1982, pp. 1992–1995.

[18] M. Airas, "Tkk aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[19] I. R. Titze and D. W. Martin, "Principles of voice production," 1998.

[20] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[21] M. Yeou and S. Maeda, "Airflow and acoustic modelling of pharyngeal," *Instrumental studies in Arabic phonetics*, vol. 319, p. 141, 2011.

[22] L. Marple, "A new autoregressive spectrum analysis algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 441–454, 1980.

[23] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–34, 2021.

[24] L. N. Smith and N. Topin, "Deep convolutional neural network design patterns," *arXiv preprint arXiv:1611.00847*, 2016.

[25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[26] U. Ha, S. Assana, and F. Adib, "Contactless seismocardiography via deep learning radars," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3372224.3419982

[27] X. Cai, Y. Cao, Y. Ren, Z. Cui, and W. Zhang, "Multi-objective evolutionary 3d face reconstruction based on improved encoder–decoder network," *Information Sciences*, vol. 581, pp. 233–248, 2021.

[28] I. Häggström, C. R. Schmidtlein, G. Campanella, and T. J. Fuchs, "Deeppet: A deep encoder–decoder network for directly solving the pet image reconstruction inverse problem," *Medical image analysis*, vol. 54, pp. 253–262, 2019.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[30] M. J. Sandage, L. W. Plexico, and A. Schiwitz, "Clinical utility of cape-v sentences for determination of speaking fundamental frequency," *Journal of Voice*, vol. 29, no. 4, pp. 441–445, 2015.

[31] C. G. Clopper and D. B. Pisoni, "The nationwide speech project: A new corpus of american english dialects," *Speech communication*, vol. 48, no. 6, pp. 633–644, 2006.

[32] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, "Shimmer™–a wireless sensor platform for noninvasive biomedical research," *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1527–1534, 2010.

[33] J. M. Levy, E. Mesel, and A. M. Rudolph, "Unequal right and left ventricular ejection with ectopic beats," *American Journal of Physiology-Legacy Content*, vol. 203, no. 6, pp. 1141–1144, 1962.

[34] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ecg analysis," *Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[35] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[36] K. Tripatbi, "Respiration and heart rate variability: A review with special reference to its application in aerospace medicine," *Indian Journal of Aerospace Medicine*, vol. 48, no. 1, pp. 64–75, 2004.

[37] L. Du, "How much deep learning does neural style transfer really need? an ablation study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3150–3159.

[38] "Uma-8 usb mic array - v2.0." [Online]. Available: https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array

[39] P. Sampson. (2021, Dec) How audacity noise reduction works. [Online]. Available: https://wiki.audacityteam.org/wiki/How_Audacity_Noise_Reduction_Works#artifacts

[40] A. Gilloire and M. Vetterli, "Adaptive filtering in sub-bands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE transactions on signal processing*, vol. 40, no. ARTICLE, pp. 1862–1875, 1992.

[41] A. Jain, B. Klare, and A. Ross, "Guidelines for best practices in biometrics research," in *2015 International Conference on Biometrics (ICB)*. IEEE, 2015, pp. 541–545.

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

[43] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and tv programs," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6209–6213.

[44] S. Zhou, J. Jia, Z. Wu, Z. Yang, Y. Wang, W. Chen, F. Meng, S. Huang, J. Shen, and X. Wang, "Inferring emotion from large-scale internet voice data: A semi-supervised curriculum augmentation based deep learning approach," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021, pp. 2–9.

[45] S. N. Ray, M. Wu, A. Raju, P. Ghahremani, R. Bilgi, M. Rao, H. Arsikere, A. Rastrow, A. Stolcke, and J. Droppo, "Listen with intent: Improving speech recognition with audio-to-intent front-end," *arXiv preprint arXiv:2105.07071*, 2021.

[46] H. Zhang, C. Song, A. Wang, C. Xu, D. Li, and W. Xu, "Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds," in *The 25th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3300061.3300125

[47] B. Stasak, Z. Huang, S. Razavi, D. Joachim, and J. Epps, "Automatic detection of covid-19 based on short-duration acoustic smartphone speech analysis," *Journal of Healthcare Informatics Research*, vol. 5, no. 2, pp. 201–217, 2021.

[48] B. Schuller, F. Friedmann, and F. Eyben, "The munich biovoice corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 1506–1510.

[49] J. Smith, A. Tsiartas, E. Shriberg, A. Kathol, A. Willoughby, and M. de Zambotti, "Analysis and prediction of heart rate using speech features from natural speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 989–993.

[50] J. Park, H. Cho, R. K. Balan, and J. Ko, "Heartquake: Accurate low-cost non-invasive ecg monitoring using bed-mounted geophones," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, sep 2020. [Online]. Available:

https://doi.org/10.1145/3411843

[51] Z. Jia, A. Bonde, S. Li, C. Xu, J. Wang, Y. Zhang, R. E. Howard, and P. Zhang, "Monitoring a person's heart rate and respiratory rate on a shared bed using geophones," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, 2017, pp. 1–14.

[52] A. Bonde, S. Pan, Z. Jia, Y. Zhang, H. Y. Noh, and P. Zhang, "Vvrrm: Vehicular vibration-based heart rr-interval monitoring system," in *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, 2018, pp. 37–42.

[53] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2015, pp. 267–276.

[54] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2016, pp. 211–220.

[55] R. Zhao, D. Wang, Q. Zhang, H. Chen, and A. Huang, "Crh: A contactless respiration and heartbeat monitoring system with cots rfid tags," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1–9.

[56] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 315–328.

[57] C. Xu, H. Li, Z. Li, H. Zhang, A. S. Rathore, X. Chen, K. Wang, M.-c. Huang, and W. Xu, "Cardiacwave: A mmwave-based scheme of non-contact and high-definition heart activity computing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–26, 2021.

[58] P. V. Rouast, M. T. Adam, R. Chiong, D. Cornforth, and E. Lux, "Remote heart rate measurement using low-cost rgb face video: a technical literature review," *Frontiers of Computer Science*, vol. 12, no. 5, pp. 858–872, 2018.