Design of a Bluetooth Earbud Heartrate and Emotion Detection System based on Multimodal Deep Learning

Zishan Wang[a]

Beijing National Day School, Beijing, China

[a]wangzishanwendy@126.com

Abstract

As the arrival of digital age, the social context becomes more competitive. People are living, working and studying under increasingly tremendous pressure. In today's society's evaluation system with the immense academic stress, anxiety, mental breakout, depression and other mental issues emerge among teenagers. The trend of the aging of population leads the elderlies to face health related challenges like loneliness, decline of cognitive competence, sleep disorders, etc. Therefore, the monitor of mental and physical health has become one of the most important issues the society need to concern. In recent years, smart watches, wristband and other wearable devices have already made some progress in the aspect of monitoring physiological signals. Still, these devices have the limitations such as low wearing comfort, low long-term compliance, insufficient privacy and security. Thus, current solutions hardly achieve a continuous, private, and highly accurate monitor on the users' mental and physical state of health. Targeting on the issues above, this research proposes a Bluetooth Earbud Heartrate and Emotion (BEHE) detection system based on multimodal deep learning. Utilizing the advantageous characteristic of portability, high privacy and convenience of earbud, we integrate photoplethysmography (PPG) sensor, physiological acoustics and inertial measurement unit (IMU) sensors to achieve the real-time heartrate and emotion detection, and provide further individualized intervention, like emotion-based music suggestion, voice assistance and feedback for emotional support or adjustment. This research provides new technical opportunity for mental health assistance.

Keywords

Bluetooth earbuds, multimodal deep learning, heart rate monitoring, emotion recognition, physiological signal processing.

1. Introduction

1.1. Research Background and Aim

As the progress of digitalization and technological development accelerates, the health state of teenagers and the elderly are of high concern. Under the influence of today's highly competitive studying environment and mental pressure, teenagers frequently face heartrate anomaly, anxiety, and emotional fluctuation. Similarly, the elderly are threatened by health risks like low mood or sleep disorders due to the decline in somatic function and reduced social interactions. Therefore, keeping track of and monitoring real-time heartrate and emotion is helpful in detecting potential health issues and providing effective interventions, significantly influencing people's daily lives.

Meanwhile, wearable devices play more significant role in health monitoring. Although smart watches and wristband already possess the function of heart rate detection, their manner of wearing might affect their comfort and privacy, and it is hard for them to achieve long-term continuous monitor. On the other hand, due to its properties like portability, lightweight,

universality, Bluetooth earbud gradually becomes the ideal physiological and emotional signal detection platform [1,2,5]. In daily scenarios like listening to music, studying, commuting etc., earbuds possess high user compliance and are appropriate to be worn for the whole day, providing a stable condition for continuous collection of human physiological data. Therefore, developing health monitor system based on earbuds is highly viable and accords with the future trend of smart devices.

Under such context, this research attempts to assemble the function of heartrate and emotion detection into Bluetooth earbuds. Specifically, it integrates multimodal sensors to capture the change in blood volume, speech signal and the micromovement of facial muscles, and utilize deep learning algorithm to construct modeling analysis, reaching the goal of real-time recognition of users' heartrate and emotion. When the system detects the users' anomaly heartrate or fluctuating emotion, it can assist to employ adjustment through playing calming music and speech feedback, making the health supervision smarter and more individualized.

1.2. A Survey on Existing Research

1.2.1. Status Quo of Heartrate Estimation Technology in Ear Devices

Heartrate is a crucial indicator for assessing human physical fitness level and cardiovascular health, related with various heart disease. Therefore, accurate monitor of heartrate is significant in managing health condition and preventing potential risks. Currently, electrocardiograph (ECG) is the most direct and accurate clinical way of reflecting cardiac electrical activity. However, ECH capture is based upon multiple electrodes attached to human body and skin. Wearing these electrodes is cumbersome, making it inappropriate to conduct enduring and continuous health monitor in daily situations.

To enhance the convenience and wearability, some researchers attempted to integrate heartrate detection function into ear devices[2,6]. For instance, some assembled bone conducting microphone into the earphone and uses the enclosed chamber formed between ear canal and tympanic membrane to amplify low-frequency heart sound signal, which is called as the Occlusion Effect[3,6]. Under quiescent condition, this method can extract clear heart sound signal, but in dynamic situation like walking or running, the noise of footstep and vibration of the surrounding air will significantly interfere the collection of heart sound, leading to inaccurate extraction of heartbeat information and reducing the precision of heartrate monitor.

At present, most of the wearable devices, including sports watches and some Bluetooth earbuds, use photoplethysmography (PPG) to monitor heartrate. PPG is a non-invasive optical measurement method, indirectly reflecting heart activity through detecting the periodic variation of subcutaneous blood volume due to heartbeat. Among the PPG signal, each pulse peak value corresponds to a heartbeat. Thereby the interval between peaks can be used to estimate heartrate and heartrate variation (HRV).

However, when measuring under motion, PPG signal is susceptible to disturbances like motion artifact, skin contact stability and temperature variation, leading to the distortion of pulse wave signal. Meanwhile, since PPG largely depends on light propagation, its detection result is also sensitive to environmental light intensity and the variation of light direction. All the factors above will lower the accuracy and stability of PPG signal detection under motional condition.

1.2.2. Status Quo of Emotion Recognition Technology in Ear Devices

Integrating emotion recognition function in wearable devices is beneficial in improving users' psychological health monitor. Earbuds have the advantage of portability, convenience, high privacy and interactivity. It can directly collect multimodal signal related to emotion and provide prompt feedback based on users' emotional state. Thus, automatic emotion detection through ear related wearable device is prospective.

Up till now, the research orientation in the field of earbud emotion recognition can be sorted into the following three main aspects:

(1) Emotion recognition based on voice tone

This method uses microphone to collect users' voice signal and detect emotion through tone, speed and energy of the audio speech signal. However, when the user is doing sports, audio signal collection may easily be interfered by noises of wind, friction and the motion artifact due to body movement, affecting the accuracy and stability of detection.

(2) Emotion recognition based on change in propagation of ultrasonic waves

Emitting ultrasonic wave from one side of the earphone and transmitting through facial tissues and finally received by the other side of the earphone, this method deduces facial micro expression via the change of propagation route of the ultrasonic wave signal. This method is based upon Acoustic Leakage and is highly sensitive to the way of wearing: it is commonly used in headphone while unachievable for earbuds that have high leak proofness.

(3) Emotion recognition by tracing facial micromotion

This method integrates inertia measurement unit (IMU) sensor in the earbuds and analyze facial expression through detecting acceleration and angular acceleration of facial muscles near the ears. Nonetheless, IMU signal is susceptible to the motion of the entire head, resulting in the confusion of emotional-related characteristics with unrelated movements [4].

Current emotion detection methods in ear-worn devices are typically monomodal, and performs unstably with the disturbance of noise, wearing manner, and change in environmental condition when the user is under motion. Therefore, how to enhance the precision of emotion detection under complex daily scenarios combining multimodal methods from facial movement to audio characteristic remains to be the main challenging of our current research.

1.3. Research Challenges

(1) Instability resulted from motion artifact

Earbuds are worn near the ear canal, susceptible to motion like walking, running or nodding. Both PPG signal, heart sound signal and IMU signal will be affected by motion artifact. Extracting relevant characteristics of heartrate and emotion under intensive disruption is the key of achieving system reliability.

(2) Integration of multimodal signal

Heartrate detection depends on PPG and heart sound, while emotion recognition requires voice signal and facial micromovement. There is great disparity in collection method and physical characteristics in different in terms of different signals. Integrating these various characteristics while avoiding intervention of unrelated information so that the model can accurately learn the correlation between multimodal signal is a crucial aspect of the research.

(3) Balance between real-time monitor and wearing comfort

Earbuds must continuously collect physiological and emotional data without affecting the users' normal daily usage and activity. Therefore, it is required to take into consideration of computational efficiency, hardware dissipation, wearing comfort, privacy and many other factors. Realizing real-time computation and stable output under the hardware constraint of existing earbuds raise higher demands and requirement to the system design.

2. Premier

2.1. Theoretical Basics of Heartrate Measurement

ECG is the most direct and accurate physiological signal of reflecting electrical activity of the heart. It is collected via attaching multiple electrodes on skin surface. Typical ECG wave form

includes P wave, T wave, and QRS complex wave, where the peak of R wave represents the moment of ventricular depolarization. The time interval between two adjacent peaks of R wave (RR interval) reflects the heartrate variation and is a core parameter of calculating HRV. Still, the collection of ECG depends on the stable contact of multiple electrodes with the skin, which is too complicated to operate with respect to daily scenarios and long-term continuous monitor. Photoplethysmography (PPG) is a non-invasive optical method for detection. Since PPG sensors have simple structure and are easy to assemble, they have been widely applied in smart watches and fitness monitor devices. PPG sensors, utilizing the fact that the amount of light absorption will vary as the blood volume changes due to heartbeat, send light of specific wavelength (usually green light) and estimate heartrate through detected reflection or the periodic change of intensity of transmitted beam. In contrast to ECG, PPG doesn't require specified gesture to achieve the measurement, which makes it appropriate for continuous tracing under motion condition. However, PPG signal is susceptible to color and thickness of skin, environmental light, and motion artifact, leading to relatively low signal to noise ratio and limiting the precision of heartrate estimation in dynamic scenario.

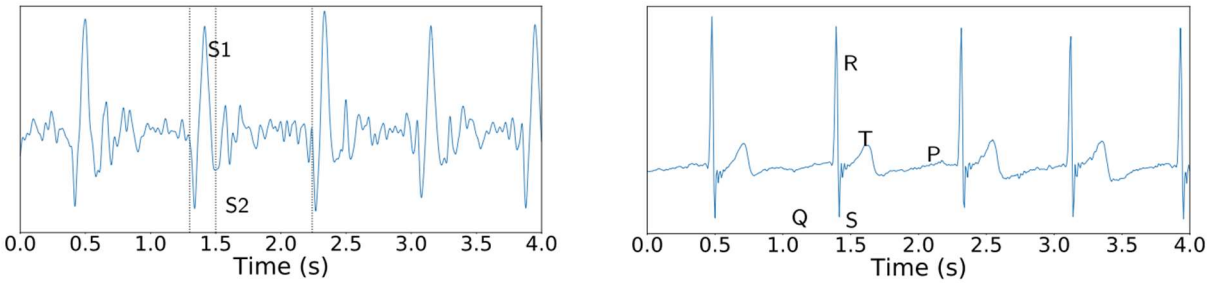


Figure 1. The left figure is heart sound signal collected by bone conduction microphone, where the first heart sound (S1) and second heart sound (S2) are marked. The right figure is the ECG signal in the corresponding period, demonstrating main heartbeat components like QRS complex wave (ventricular depolarization), T wave (ventricular repolarization) and P wave (atrial depolarization)[7].

To overcome the limitations mentioned above, this research introduces the bone conduction signal as an assisting signal detection mode. Bone conduction microphone (BCM) can collect low-frequency heart sound signal induced by heartbeat and breathing, and is insusceptible to light condition or surface sliding, possessing the relatively high stability for detection under motion. Therefore, fusing the BCM signal and PPG signal through a multimodal method allows us to reconstruct a simulation of the ECG wave signal, thereby enhancing the robustness of heartrate estimation. The left side of Figure 1 shows the heart sound signal collected by in-ear bone conduction microphone, while the right side is the ECG signal in the same period of time. It is obvious that the two have a satisfactory correspondence in terms of the peak value of the waveform in the heart's rhythm, verifying the feasibility of detecting heartrate using BCM signal.

2.2. Theoretical Basics of Emotion Recognition

Firstly, the internal microphone of the earbuds can collect users' voice data. The voice not only includes the content of the speech but also contains emotion related voice characteristics, like tone, speed and change in energy. Previous study by other researchers has proved that the speeches of same content but under different emotion display disparate acoustic manifestation, so we can deduce the emotion of the users by analyzing the prosodic features of their speech. This study utilizes large-scale pre-trained speech models to extract speech features and

separates the speech embedding into two parts: content features and prosodic features, enhancing the robustness of the precision of the recognition.

Secondly, the facial expression is closely related to emotion as well. To capture the movement of facial muscles, this study integrates multi-axis IMU sensors in the earbuds to detect the micromovement near the ears and on the face. Facial Action Coding System (FACS) is a structured description system for facial movements composed of action units that can be used to construct a corresponding relationship between the variation of facial expression and the variation of the muscle's movements. Based on the standard of FACS, we collected and analyzed the IMU signals that are in correspondence to different action units.

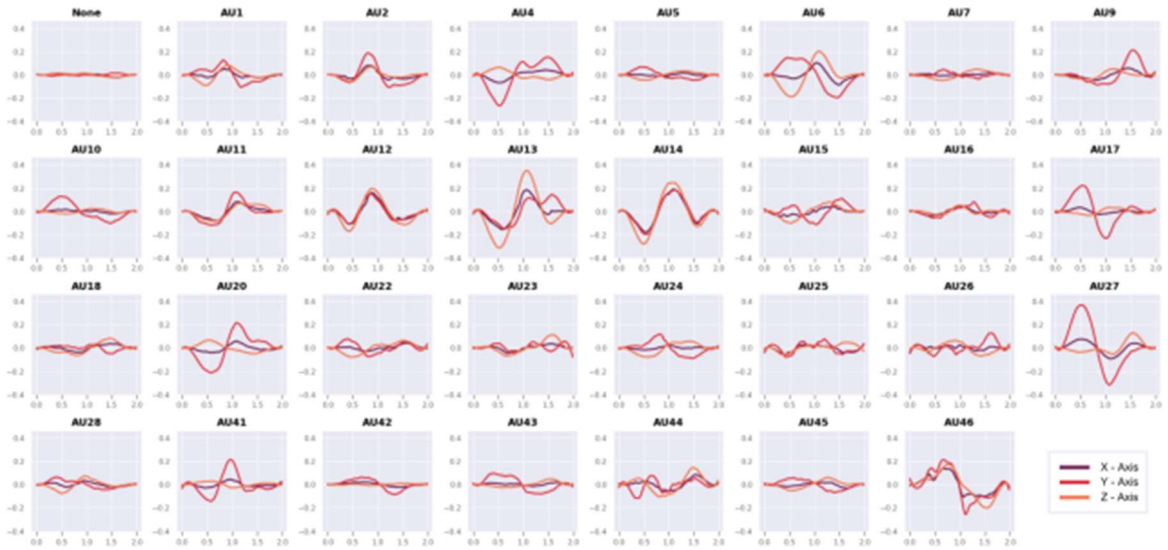


Figure 2. The schematic diagrams illustrating the changes in the three-axis acceleration signals corresponding to different facial action units (Action Units, AUs) collected by the built-in IMU sensor in the earbuds. Each AU graph shows the time-domain IMU signal in X, Y, Z axis. The difference of muscle action triggered by different facial expression can be demonstrated through the change in waveform, providing effective evidence for emotion recognition[8].

Existing research results shows that there are constant discrepancies on the movement pattern of muscles when people show different facial expression. For instance, in the IMU signal corresponding to AU10 (Lip stretcher) and AU16 (Lower Lip Depressor), we can observe the characteristic change in different directions, showing that IMU can effectively reflect the information of muscle movement related to emotion, which provides reliable support to the emotion detection.

3. Methods

3.1. Overall Methods and Framework

This research proposes a multimodal health monitor system based upon Bluetooth earbuds, achieving the real-time estimation and detection of heartrate, heartrate variation, and emotional state via the integrated PPG signal, BCM signal, voice signal and facial micromovement signal. The overall framework of the system has two main modules: heartrate detection module and the emotion recognition module (Fig.3).

In the heartrate detection part, the integrated PPG sensor and bone conduction microphone collect change in blood volume and heart sound signal respectively. The collected signal will undergo signal preprocessing procedures including frame segmentation, band-pass filtering etc.

to eliminate noise and motion artifact, and then we extract effective features of the signal via time-frequency conversion (Mel spectrogram). Afterwards, we use U-Net neuro network for fusion learning of multimodal features, reconstructing the waveform approximating ECG and thereby realizing the highly precise estimation of heartrate and heartrate variation.

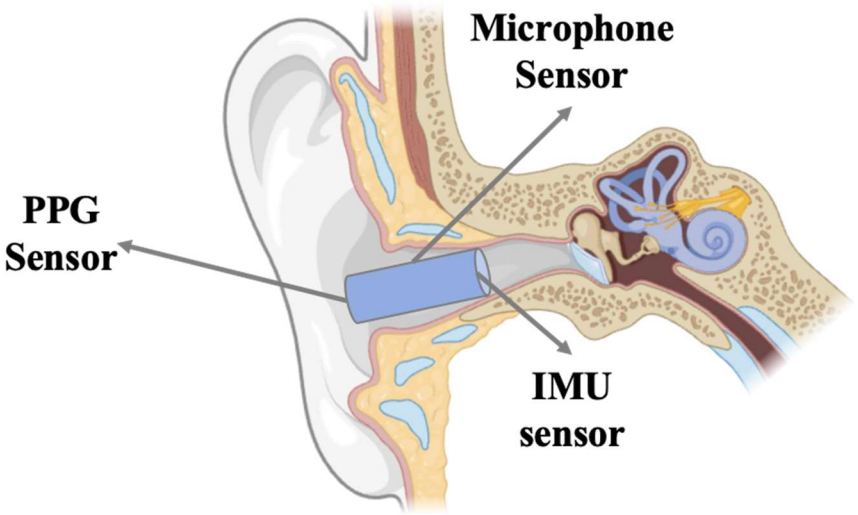


Figure 3. Partial diagram of multimodal sensors of the earbud. The earbud integrates PPG sensor, bone conduction microphone, and IMU sensor to respectively capture variation of blood volume, heart sound signal and micromovement feature of facial muscles, providing support of data for heartrate and emotion detection.

In the emotion recognition aspect, the internal IMU sensor of the earbuds record real-time acceleration caused by the micromovement of muscles near users' ears and faces while the microphone captures the users' voice signal. After preprocessing, the voice signal is inputted into cloud-based voice model, acquiring the embedded voice features that are decoupled to two emotion-related factors: semantic content feature and prosodic feature. Finally, fusing the vocal features and IMU motion features together in the fully connected neural network, we get to recognize and detect users' real-time emotional state.

3.2. Signal Preprocessing

In this research, we use bone conduction microphone to collect heart sound signal and use PPG sensor to capture the signal of variation in blood volume. Since the raw data contains immense environmental noise and motion artifact, we need the signal to be systematically preprocessed to enhance the precision of the later recognition of the model. First, we collect both data at a sampling rate of 1000Hz, meaning that we capture 1000 data points per second. In terms of frame segmentation, we use a window length of 2 seconds to cut the continuous signal and set 1.5 seconds of overlap of adjacent windows, ensuring the continuity of temporal sequence information and the integrity of cardiac rhythm characteristics.

To inhibit the disruption of unrelated signals, the heart sound signal undergoes band-pass filtering methods, which filter out the high-frequency disturbance, like background noise, sound produced by friction between ears and earbuds, leaked sound of music played in the earbuds etc., and keeps the component of low-frequency heart sound, making the main peak structure of the signal more prominent. Similarly, PPG signal also undergoes the basic noise cancellation processing to minimize the effect of motion artifact on the pulse waveform.

Next, we apply short-time Fourier Transform (STFT) to the two filtered signals and utilize Hanning window to reduce leak of low frequency, thereby converting time-domain signal to time-frequency domain signal. To facilitate the extraction of physiological features by the deep learning model, we furthermore convert the spectrum expression to 64 logarithmic Mel-frequency spectrograms of Mel-band. Each 2-second window ultimately generates a 64×64 characteristic matrix, containing both temporal dimension and important information of frequency.

Since both sides of the earbuds are equipped with BCM and PPG sensors, the system convert four-way signal into Mel spectral feature each time and accumulate in the vertical direction and getting an input feature tensor with a size of $64 \times 64 \times 4$. This multimodal feature representation method is able to unify the signal of heart sound and blood volume variation into the same representation space, lying a foundation for later ECG reconstruction and heartrate estimation task based on deep learning.

3.3. Network Structure

In this study, we didn't let the deep learning model directly output a heartrate value; instead, we start from reconstruction of ECG waveform from multimodal in-ear signals and then calculate the heartrate based on the simulated ECG waveform. The reason why we do this is that in comparison with learning the complex nonlinear relationship between heart sound and heartrate, it is clearer and more stable to learn the correspondence between heart sound and ECG waveform; additionally, the latter method provides a greater possibility for the calculation of more diverse physiological indicators, such as heart rate variation.

The input for the deep learning network is the four-channel logarithmic Mel spectrogram, we acquired after signal processing in the previous section, whose dimension is $64 \times 64 \times 4$. It includes a heart sound signal and PPG signal for each ear. The output shape of the network is $64 \times 64 \times 1$, corresponding to the reconstructed mono-channeled ECG feature graph.

This study utilizes U-Net as signal reconstruction network. U-Net is constructed by a symmetric structure formed by Encoder and Decoder. Encoder extract multidimensional features through multi-layer convolution and Rectified Linear Unit (ReLU) activation function and gradually compress the feature diagram under the max-pooling operation, enabling the model to learn deep-layer modes of relevant heartbeat rhythm. The decoder part performs layer-by-layer upsampling of the features through deconvolution so that the output restores to a same time-frequency resolution as the input. Meanwhile, the feature of the corresponding layer of the encoder will fuse with the decoder through skip connection to maintain the partial details and key waveform structure, which is especially important to the restoration of the highly mutating features of the R-wave peak.

Assume that the deep learning model is f(.), the input signal is X, the reconstructed ECG signal outputted by the network is $\widehat{\mathrm{Y}}$. Therefore, the prediction process of the network can be expressed as:

$$\widehat{\mathrm{Y}} = f(X;\ \theta)$$

To make the reconstructed ECG closer to the actual ECG signal, this research applies the commonly used Mean-Square Error (MSE) as the loss function:

$$\mathcal{L}_{MSE} = (\widehat{\mathrm{Y}} - \mathrm{Y})^2$$

Where Y represents the actual ECG signal and $\hat{Y}$ represent the model's predicted signal. Through iterative optimal, the predicted waveform gradually approximates the actual waveform, acquiring a high-quality ECG reconstruction result.

The reconstructed ECG signal contains the typical P wave, QRS complex wave and T wave, where the peak of R wave represents the instantaneous time at which the ventricular systole and the intervals between the two adjacent R wave (RR interval) directly related to heartrate. Therefore, we extract the positions of all the peaks of the R wave and estimate heartrate according to the time difference:

$$HR = t_{R_{i+1}} - t_{R_i}$$

Where $t_{R_{i+1}}$ and t_{R_i} is the time at which the two consecutive peaks of R wave appear. Through sliding the window, we can calculate the average RR interval, and in turn, getting a stable estimation of heartrate.

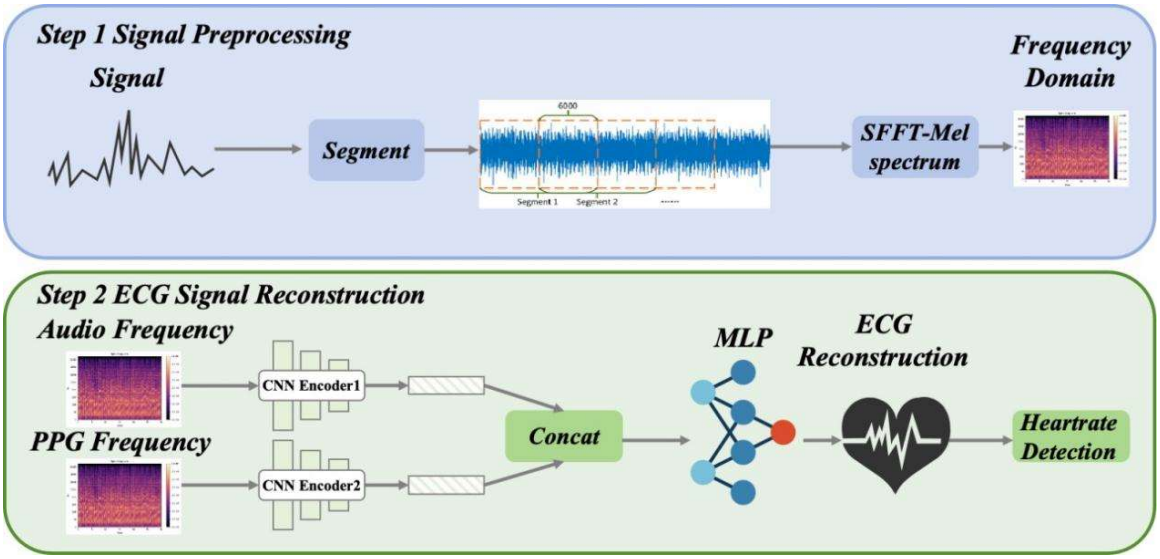


Figure 4. Flow diagram of heartrate estimation method based on multimodal in-ear signal

By the procedures above, this research realizes a complete process from in-ear multimodal signal→ECG reconstruction→heartrate estimation, enormously boosting the accuracy of heartrate monitor of ear wearing devices under dynamic and in-motion scenarios.

In the emotion recognition module, this research utilizes in-ear microphone and IMU sensors to simultaneously collect voice signal and micromovement signal of facial muscle, estimating users' emotional state through multimodal fusion method.

The voice signal first undergoes preprocessing of noise cancellation and framing, and we extract tone-sensitive acoustic features. For example, parameters like speed of speech, energy variation and the fundamental frequency outline all reflect users' sense of strain, exhaustion, delight, depression etc. Later, preprocessed voice signal is inputted to the deep neural network that have been pretrained through a large scale of vocal data, extracting emotion-related embedded features. To prevent the semantic content of the speech from influencing the judgement of emotion, we decouple the embedded voice into two parts: semantic feature and prosodic feature, and retain the prosodic feature for later recognition, allowing the model to focus on the emotional expression itself.

In addition, the IMU sensor in the earbuds can record acceleration of facial muscles due to micromovement. According to FACS, different emotion triggers different muscular contraction patterns, like actions of lowering the eyebrows, depressing the lips, etc. By extracting three-axes acceleration signal as features and analyzing its dynamic development as time varies, we can capture the information of users' micro facial expression.

In the fusion phase, we joint the vocal prosodic feature and IMU facial movement feature together in a unified feature space and import them into fully connected neural network for the classification task. The network ultimately output a probability distribution of the emotion classification, thereby completing the real-time judgement of users' emotion state. This multimodal fusion strategy remedies the drawback of susceptibility to environmental disturbance of monomodal feature, enhancing the performance and stability of recognition task in daily scenarios.

4. Experimental Process

To test the effectiveness of the earbud multimodal heartrate and emotion detection system proposed by this study, we separately tested the heartrate estimation module and the emotion recognition module through experiment. In the experiment, we use earbuds to collect bone conduction heart sound signal, PPG signal, voice signal and the muscular micromotion signal simultaneously. All the subjects are tested under normal breathing and natural conditions to simulate the realistic, continuous using scenario in daily life.

In heartrate estimation, the subjects simultaneously use ethe standardized clinic method to collect actual ECG signal as a control group. The experiment uses data collected in various time period, including stationary sitting position, standing, slow walking etc. to estimate the model's adaptability to variation of heartrate under various scenarios in motion. The collected multimodal signal undergoes preprocessing and is then imported into U-Net model to reconstruct ECG waveform. The positions of R peaks of the new ECG signal are extracted for heartrate calculation. Comparing the reconstructed ECG with the actual signal, we can evaluate how accurate the model is in estimating heartbeat rhythm restoration.

In the emotion recognition experiment, the subjects completed multiple emotional expression under the commands, including emotional state like joyful, calm, nervous, depressed. When executing the task, the subjects also need to say a sentence in a prosodic feature that can express its emotion, while the IMU sensor record facial muscle motion. All the signals are imported into classification neural network for feature processing, and through the probability distribution of the classification of emotion outputted by the model, we can determine its correctness comparing with the actual emotional expression delivered by the subject. To ensure the objectiveness of the result, this experiment repeats the measurement multiple times and the results are compared for consistency with the real emotional labels that were manually annotated.

5. Conclusion

This study proposes a heartrate and emotion detection system based on multimodal in-ear sensors, combining bone conduction heart sound signal, PPG signal, voice information and facial muscle micromotion and thereby realizing the continuous recognition and smart analysis of the users' physiological and emotional state. In the heartrate detection aspect, reconstructing ECG signal using multimodal fused features and U-Net effectively improved the stability and accuracy of heartrate estimation under dynamic scenario; in the emotion recognition aspect, using voice prosodic features and the facial micromovement information captured by IMU sensor, it successively achieves automatic basic emotion classification and recognition.

References

[1] Islam, M. S., Rahman, M. M., Bin Morshed, M., et al. (2025). BallistoBud: Heart rate variability monitoring using earbud accelerometry for stress assessment. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–21).

[2] Fu, Y., Sun, K., Wang, R., et al. (2025). Enabling cardiac monitoring using in-ear ballistocardiogram on COTS wireless earbuds [Preprint]. arXiv. https://arxiv.org/abs/2501.06744

[3] Cao, Y., Cai, C., Li, F., et al. (2024). Enabling passive user authentication via heart sounds on in-ear microphones. *IEEE Transactions on Dependable and Secure Computing*. Advance online publication.

[4] Zhang, S., Lu, T., Zhou, H., et al. (2023). I am an earphone and I can hear my user's face: Facial landmark tracking using smart earphones. *ACM Transactions on Internet of Things*, 5(1), 1–29.

[5] Zou, Y., Weng, J., Lei, H., et al. (2024). EarPrint: Earphone-based implicit user authentication with behavioral and physiological acoustics. *IEEE Internet of Things Journal*, 11(19), 31128–31143.

[6] Chen, T., Yang, Y., Fan, X., et al. (2024). Exploring the feasibility of remote cardiac auscultation using earphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (pp. 357–372).

[7] Butkow, K. J., Dang, T., Ferlini, A., et al. (2024). An evaluation of heart rate monitoring with in-ear microphones under motion. *Pervasive and Mobile Computing*, 100, 101913.

[8] Li, Y., Zhang, T., Zeng, X., et al. (2024). Auglasses: Continuous action unit based facial reconstruction with low-power IMUs on smart glasses [Preprint]. arXiv. https://arxiv.org/abs/2405.13289