PAN
ARCHIVES OF ACOUSTICS
Vol. **46**, No. 1, pp. 41–53 (2021)
DOI: 10.24425/aoa.2021.136559

Research Paper

Heart Rate Detection and Classification from Speech Spectral Features Using Machine Learning

Mohammed USMAN[(1)*], Mohammed ZUBAIR[(1)], Zeeshan AHMAD[(1)], Monji ZAIDI[(1)], Thafasal IJYAS[(1)], Muneer PARAYANGAT[(1)], Mohd WAJID[(2)], Mohammad SHIBLEE[(3)], Syed Jaffar ALI[(4)]

[(1)] *Department of Electrical Engineering*
King Khalid University
Abha, 61411, Saudi Arabia
[*]Corresponding Author e-mail: omfarooq@kku.edu.sa

[(2)] *Department of Electronics Engineering*
Aligarh Muslim University
Aligarh, 202001, India

[(3)] *Department of Computer Engineering*
Taif University
Taif, 21944, Saudi Arabia

[(4)] *Department of Computer Engineering*
King Khalid University
Abha, 61411, Saudi Arabia

(*received May 20, 2020; accepted October 5, 2020*)

Measurement of vital signs of the human body such as heart rate, blood pressure, body temperature and respiratory rate is an important part of diagnosing medical conditions and these are usually measured using medical equipment. In this paper, we propose to estimate an important vital sign – heart rate from speech signals using machine learning algorithms. Existing literature, observation and experience suggest the existence of a correlation between speech characteristics and physiological, psychological as well as emotional conditions. In this work, we estimate the heart rate of individuals by applying machine learning based regression algorithms to Mel frequency cepstrum coefficients, which represent speech features in the spectral domain as well as the temporal variation of spectral features. The estimated heart rate is compared with actual measurement made using a conventional medical device at the time of recording speech. We obtain estimation accuracy close to 94% between the estimated and actual measured heart rate values. Binary classification of heart rate as 'normal' or 'abnormal' is also achieved with 100% accuracy. A comparison of machine learning algorithms in terms of heart rate estimation and classification accuracy is also presented. Heart rate measurement using speech has applications in remote monitoring of patients, professional athletes and can facilitate telemedicine.

Keywords: heart rate from speech; machine learning; MFCC; regression and classification; speech as a biomedical signal.

1. Introduction

Vital signs of human body are conventionally measured using medical equipment. These can be complicated to use, expensive and also cause inconvenience to the patient/individual. This is especially true for measuring vital signs of athletes during their training which involves intense physical activity. Connecting electrodes, sensors and other medical equipment on athletes while they are training is likely to be intrusive and affect their performance. Physiological as well as emotional changes in an individual result in variations in the speech produced (Trouvain, Truong, 2015; Science Encyclopedia, 2019; Borkovec *et al.*,

1974; Ramig, 1983; Reynolds, Paivio, 1968). Ageing, health condition, stress level, exposure to pollution as well as physical exercise and activity are some factors which can cause physiological changes in the human body. While existing literature suggests that speech production process is affected by physiological changes in individuals, the effect of such physiological changes on the actual speech parameters needs thorough investigation (Trouvain, Truong, 2015). The work presented in this paper is directed towards the estimation of heart rate from features extracted from speech signals using machine learning. While there is sufficient evidence from published literature linking physiological and emotional conditions to speech production, research on the actual estimation of physiological parameters accurately using speech is still at a nascent stage. If the prediction is indeed accurate, it would substantiate clinical findings that there exists a correlation between physiological condition and speech characteristics of individuals and pave the way for non-invasive and non-contact based, remote medical monitoring. It should be noted however that such speech based medical monitoring shall complement existing medical devices rather than replace them.

The topic of this research has the potential for rapid development leading to a plethora of application scenarios, if estimation accuracy is improved. The results presented in this article will have valuable impact and can lead to interdisciplinary research involving electronics, signal processing and medicine. By being able to measure vital parameters of the human body without complex and expensive medical equipment, it will simplify the cost of medical diagnosis and treatment. Furthermore, it can make it possible for medical practitioners and professional sport trainers to monitor patients/athletes from remote location by collecting their speech samples over the telephone or internet.

2. Speech as a biomedical signal

2.1. Related work

The existence of relationship between human speech and physiological parameters is evident from published literature. In (Schuller *et al.*, 2013), measurement of heart rate and skin conductance as well as classification of pulse rate as 'high' or 'low' has been done using audio recordings of breath and sustained vowel sounds with nominal accuracy. Extraction of electrocardiogram (ECG) features from two dimensional spectrum of vowel speech is demonstrated in (Skopin, Baglikov, 2009; Mesleh *et al.*, 2012) in which the vowel sound 'i' as in the word 'email' is shown to yield better accuracy compared to the other vowel sounds. Heart rate extraction using statistical analysis of speech is presented in (Kaur, Kaur, 2014), but there is no mention regarding the accuracy of the technique. A data mining approach is used in (Sakai, 2015b) to establish a correlation between heart rate and vocal frequency from which heart rate is estimated using multiple speech recordings from only two users. A comparison of different classifiers to detect emotions based on Mel frequency cepstrum coefficients (MFCC) is presented in (James, 2015). Blood pressure (BP) detection from speech using support vector machine (SVM) is suggested to be feasible in (Sakai, 2015a) with high correlation between estimated and actually measured values of BP. It is also shown in (Orlikoff, Baken, 1989) that heartbeat has an influence on the vocal fundamental frequency causing it to fluctuate. In (Schuller *et al.*, 2014), measurement of heart rate and skin conductance from various speech features, using machine learning algorithms such as support vector regression (SVR), SVM, artificial neural networks (ANN) as well as linear regression has been presented with moderate accuracy and the authors conclude that MFCC features are particularly relevant for the task of measuring heart rate from speech.

Heart rate is affected by physical activity performed by an individual and based on observation, speech is affected when performing physical activity such as exercise or sport. It has been shown in (Usman, 2017) that the accuracy of speaker recognition system based on MFCC is reduced when speech is recorded immediately after intense physical activity, suggesting that speech features are altered. Heart rate variation depends on the level of physical activity as well as the fitness of the individual, in addition to other factors (James, 2015). Furthermore, the physiological response to activity depends on the intensity, duration and regularity of performing the activity (Burton *et al.*, 2004). These strongly suggest that there exists some correlation between speech and heart rate which provides a basis and motivation for conducting this research. Accurate prediction of heart rate and other physiological parameters based on speech signals has the potential to revolutionize medical care by monitoring patients remotely and provide timely medical intervention. With the advent of telemedicine and wide availability of portable medical devices, this could be a game changer as the ubiquitous and humble smartphone can extend its functionality as a medical device, without the need to incorporate additional sensors.

Non-contact based measurement of physiological parameters such as heart rate, heart rate variability, respiratory rate and blood volume pulse, by applying independent component analysis (ICA) to facial images and video has been proposed in (Poh *et al.*, 2011). Extraction of heart rate, heart rate variability, blood oxygen saturation and breathing rate using video of finger tip has been presented in (Scully *et al.*, 2012). While there is significant evidence from literature suggesting the existence of correlation between speech and

certain physiological parameters, the focus of this work is to measure heart rate from speech and compare it with actual heart rate measured concurrently at the time of recording speech using a conventional medical device.

3. Materials and methods

3.1. Speech samples and heart rate measurement

Speech recordings have been made for 42 individuals, all male, in the age group of 20–45 years using a Logitech H540 headphone set, which is equipped with a noise-cancelling microphone to minimize background noise, in a quiet office environment. The sentence uttered is 'A quick brown fox jumped over the lazy dogs' which is chosen in order to capture the sounds of all letters in the English alphabet. The duration of each audio recording is 5 seconds stereo format and sampling rate is f_s = 16 000 samples per second, which is a standard value used in speech processing since it corresponds to wideband (8 kHz) representation of speech that faithfully restores all frequency components of the speech signal (Usman *et al.*, 2018). As most of the salient features of speech lie within the 8 kHz bandwidth, increasing the sampling rate beyond 16 000 samples per second leads to a point of diminishing returns while increasing the length of data leading to increased computational complexity. Lower sampling rate can cause aliasing effect of some high frequency components and hence 16 000 samples per second is considered a reasonable choice to avoid aliasing effects as well as avoiding unnecessary increase in complexity. Higher sampling rate is required for non-speech sounds such as breathing sounds, cough sounds etc. The focus of this article is on speech sounds and therefore 16 000 samples per second is an appropriate sampling rate. The recordings are stored on a PC as uncompressed WAV file format that uses a quantization depth of 16 bits per sample (Kabal, 2017) resulting in audio bit rate of 256 kbps. Heart rate measurements of each individual are taken using a pulse oximeter (CONTEC Model No. CMS50DL) concurrently during the speech recording. These measurements are used for comparison with the predicted heart rate values to obtain the accuracy of the machine learning methods used to predict heart rate from speech. A pulse oximeter is a medical device that is attached to the finger tip to measure pulse rate and blood oxygen saturation. "Pulse rate is exactly equal to the heart rate as the contraction of the heart leads to a noticeable pulse" (MacGill, 2017).

3.2. Speech pre-processing

The speech recordings are preprocessed to remove any unwanted components as well as silence intervals in speech that may have been introduced during the recording process. PC audio cards introduce a small DC offset (Partila *et al.*, 2012) that is removed using a DC removal filter which is a first order infinite impulse response (IIR) filter. Silence intervals in the recorded sentence, which do not contain voice activity are removed using a voice activity detection (VAD) algorithm (Tan, Lindberg, 2010). The VAD algorithm identifies speech frames containing voice activity by assigning higher frame rate to consonant sounds, lower frame rate to vowel sounds and no frames to silence intervals. The effect of noise is also mitigated by the VAD algorithm using a posteriori signal to noise ratio (SNR) weighting to emphasize reliable segments of speech even under low SNR. The identified frames are then concatenated together resulting in uttered sentence without silence intervals and improved SNR.

3.3. Feature extraction

Feature extraction is a term derived from the discipline of pattern recognition and refers to characterizing a signal in a manner that allows some algorithm to recognize a pattern (Wolf, 1980). We extend this definition to "characterize a signal that allows some algorithm to recognize a pattern or some 'intrinsic' parameter associated with that pattern". We conjecture that such an intrinsic parameter obtained from patterns in speech features to be a representation of a physiological parameter such as heart rate of the individual who uttered that speech. This is based on the fact that speech production process involves movement of air from the lungs and through the vocal tract. As lungs interact with heart for oxygenation of blood, it is suggested in (Reilly, Moore, 2003) that cardiovascular responses are affected by cognitive activity such as reading, which involves speech production. As breathing is utilized for speech production, the inhalation and exhalation rates are governed by the speech production mechanism, thus altering the breathing pattern during speech production (Von Euler, 1982). Heart rate variability due to changes in respiratory pattern, termed as respiratory sinus arrhythmia (RSA) is discussed in (Yasuma, Hayano, 2004). These strongly suggest that speech signals contain information about heart rate and perhaps other physiological parameters which may be determined by extracting appropriate speech features and processing those features using machine learning algorithms. Results presented in this article indeed validate this idea as heart rate values are obtained from speech features with a high degree of accuracy. A variety of speech feature extraction techniques are available in the literature for various speech processing applications such as speech recognition, speaker recognition, speech enhancement etc. Some of the well-known techniques are linear prediction coefficients (LPC), lin-

ear prediction cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), feature extraction based on principal component analysis (PCA) and wavelets (Magre *et al.*, 2013).

LPC, which represents speech parameters by an all pole filter using auto-regressive modeling and LPCC, which are cepstral coefficients computed from a smoothed auto-regressive power spectrum were widely used in automatic speech recognition until the introduction of MFCC (Huang *et al.*, 2001). Since the introduction of MFCC's in 1980, it has been widely used in several speech processing applications and considered to be the most popular feature extraction method. Discriminating features in speech are better represented in spectral domain and temporal variation of spectral components also have a significant effect in characterizing speech. MFCC captures the spectral domain details along with their temporal variations elegantly with a low dimensional feature vector. A detailed discussion of MFCC is available in (Davis, Mermelstein, 1980). PLP is a spectral warping technique used to model and obtain an estimate of human auditory spectrum (Hynek Hermansky, 1990) and is more suited for speech recognition application. PCA is used to reduce the dimensionality of feature vectors by transforming the feature vectors to lower dimension (Huang *et al.*, 2001). Wavelets have also been proposed in the literature to obtain a modified version of MFCC in which the discrete wavelet transform (DWT) is applied instead of discrete cosine transform (DCT) in the MFCC computation process, resulting in what is termed as Mel frequency discrete wavelet coefficients (MFDWC) (Tufekci, Gowdy, 2000). DWT has the advantage of providing better time-frequency resolution but there is not enough evidence from literature to purport a broad range of applications for DWT based speech features. Relative spectra (RASTA) is a technique which focuses on mitigating channel effects to improve speaker recognition systems. It is suggested that RASTA makes short term spectrum based techniques such as PLP more robust to linear spectral distortions (Hermansky, Morgan, 1994). In this work, MFCC's have been used as features applied to machine learning algorithms in order to estimate heart rate from speech signals. MFCC features have been chosen due to their ability to capture spectral details along with their temporal variations. Some existing results in literature to estimate heart rate from speech are also based on MFCC.

3.4. MFCC computation

Implementation of MFCC computation is available in (Davis, Mermelstein, 1980). The specifics of MFCC computation in the context of this work are described here for the sake of completeness. The preprocessed speech signal in which silence intervals are removed is framed using a Hamming window having length 256 samples with a 50% overlap with adjacent windows. At sampling rate of $f_s = 16\,000$ samples per second, this corresponds to each frame having a length of 16 ms which is within the stationary duration of 20–25 ms for speech signals and overlap duration of 8 ms. Each frame 'i' is denoted as $x_i(n)$, where $n = 1, 2, ..., 256$. An N-point Fast Fourier Transform (FFT) is computed with $N = 256$ for each 16 ms speech frame to obtain the spectrum of that segment. The combined process of windowing and FFT is represented as

$$X_i(k) = \sum_{n=1}^{N} x_i(n)h(n)\exp\left(-\frac{j2\pi kn}{K}\right) \quad \text{for} \quad k = 1, ..., K, \tag{1}$$

over the entire range of i, i.e. the total number of frames, k denotes the discrete Fourier transform (DFT) coefficients computed using FFT and $K = 256$. The energy spectral estimate of each frame is then computed as

$$E_i(k) = |X_i(k)|^2 \quad \text{for all } i. \tag{2}$$

Energy spectral estimate is used rather than power spectral estimate as the length of each speech recording is short (less than 5 seconds) and the frame duration of 16 ms is not considered infinitesimally small relative to the length of each recording. The power spectral estimate is used when the signal duration is long enough to be considered infinite relative to the frame duration over which power is computed (Oppenheim, Verghese, 2015). Mel filterbank comprising of 20 triangular filters with 50% overlap between adjacent filters is then applied to each frame. Since the sampling rate is 16 000 samples per second, the frequency range for each frame extends from zero Hz to 8000 Hz. The corresponding minimum and maximum Mel frequencies are zero Mels and 2834.99 Mels respectively obtained using (Lyons, 2012)

$$\text{Mel}(f) = 1125\ln\left(1 + \frac{f}{700}\right). \tag{3}$$

To generate a filterbank with 20 filters, 20 linearly spaced points are generated between zero and 2834.99. The resulting Mel frequencies are $\text{Mel}(f) = \{0, 135, 270, 405, 540, 675, 810, 945, 1080, 1215, 1350, 1485, 1620, 1755, 1890, 2025, 2160, 2295, 2430, 2565, 2700, 2835\}$. The first Mel window extends from zero to 270 Mels, the second Mel window from 135 to 405 Mels and so on. The conversion from 'Mels' to 'Hz' is performed using

$$f = 700\exp\left\{\left(\frac{\text{Mel}(f)}{1125}\right) - 1\right\}, \tag{4}$$

resulting in $f = \{0, 89.2, 189.9, 303.3, 431.3, 575.5, 738.1, 921.5, 1128.2, 1361.3, 1624.1, 1920.4, 2254.5, 2631.2,$

3055.9, 3534.7, 4074.6, 4683.4, 5369.8, 6143.6, 7016.2, 7999.9}. It should be noted that Mel frequencies are linearly spaced and the corresponding frequencies in Hz are logarithmically spaced. These logarithmically spaced frequencies are mapped to their nearest FFT bins as follows (Lyons, 2012)

$$\mathrm{FFT}_{\mathrm{bins}} = \lfloor (N+1) \times f/f_s \rfloor, \tag{5}$$

where $\lfloor . \rfloor$ is the floor operator, N is the number of FFT points used, and f_s is the sampling rate. For these chosen values, the FFT bin corresponding to 8000 Hz is bin 128. The sequence of Mel warped FFT bins is $\mathrm{FFT}_{\mathrm{bins}} = \{0, 1, 3, 4, 6, 9, 11, 14, 18, 21, 26, 30, 36, 42, 49, 56, 65, 75, 86, 98, 112, 128\}$.

Thus, 20 Mel filter windows are produced each having a length of 256, which is chosen to be the same as the number of FFT points computed for each frame. Each of the 20 Mel filters is multiplied with the energy spectrum $E_i(k)$ and the coefficients are added to obtain the energy within each band. For each 16 ms speech frame, this results in a vector of length 20 where each element represents the signal energy within a Mel filter band. The log-energy is computed by taking the logarithm of the 20 filter-bank energies. The log filter-bank energies so obtained have a high degree of correlation due to overlapping filters in the filter-bank.

A decorrelation transform is applied to decorrelate the Mel-spectral vector. Discrete cosine transform (DCT) is shown to be a near optimal decorrelation transform for log spectra of speech (Merhav, Lee, 1993; Logan, 2000). DCT is therefore applied to the 20 log filter-bank energies resulting in 20 coefficients for each 16 ms speech frame which are called the Mel Frequency Cepstral Coefficients. MFCC's are computed for all the frames that comprise the pre-processed speech signal resulting in a matrix of size $20 \times I$, where I is the total number of frames. This matrix of MFCC coefficients is analyzed using machine learning algorithms to estimate the heart rate of individuals from their speech signals and also to classify heart rate as 'normal' or 'abnormal'.

3.5. Predictive analytics for heart rate estimation

In this study, we have utilized the Microsoft Azure Machine Learning Studio (MAMLS) cloud platform which is accessible through a web interface. MAMLS allows for high volume secure data storage and transmission, computational analytics and remote visualization. Machine learning algorithms available in MAMLS have been tuned and configured for maximizing resting heart rate (HR) regression and classification accuracy. Machine learning techniques learn the statistical relationship between input data (e.g. MFCC coefficients extracted from speech signals) and output data (e.g. HR) by fitting a flexible model to the data. The model hyper-parameters are optimally parameterized to minimize the regression/classification error in an independent test dataset, thereby allowing for creating a generalized model that can perform well not only on the training dataset, which would give rise to an over fitted solution but also on test dataset. For comparative analysis, six state-of-the art machine learning algorithms available in MAMLS, have been considered for regression (numerical estimation of HR of the individual) and binary classification analysis. Here, we have briefly summarized them for brevity. Linear Regression (LiR) is a very common statistical method utilized in machine learning for fitting a line to the input features and measuring the error. LiR tends to work well on high dimensional data sets that lack complexity (Kutner *et al.*, 2004). Boosted Decision Tree (BDT) is classed as an ensemble learning technique, in which consecutive tree corrects for the errors of the previous tree thereby minimizing classification error. Class and value predictions are based on the entire ensemble of trees (Bühlmann, Yu, 2003). Decision Forest (DF) is another ensemble learning technique, in which each generated tree votes for the most popular class (Criminisi *et al.*, 2011).

Neural Networks (NN's) are a set of interconnected layers. A typical NN, comprises of neurons in the three layers. The input feature set forms the first layer and is linked to the output layer via several interconnected hidden layers in the middle of the network. Each neuron is responsible for processing the input variables and passes the computed values to the neuron in the subsequent layer (Zhang *et al.*, 1998). Logistic Regression (LoR) is another statistical technique for analyzing data in which there are one or more independent variables that determine an outcome. The outcome is normally measured with a dichotomous variable (i.e. having only two possible outcomes) (Dreiseitl, Ohno-Machado, 2002).

Support Vector Machines (SVM) work on the basic principle of recognizing patterns in a multi-dimensional hyper-plane and estimating a maximum margin between samples of the binary classes in a multi-dimensional input feature space (Nasrabadi, 2007). All of the algorithms mentioned above, have been successfully used in various application domains due to their relatively fast training, excellent performance and their robustness to over fitting. The performance of the various proposed regression and binary classification models are evaluated based on the metrics listed in Table 1. Depending on the task, the listed evaluation metrics (Roychowdhury, Bihis, 2016) for regression or binary classification are used.

Table 1. List of performance evaluation metrics.

Evaluation metric	Definition
Regression analysis	
Mean absolute error	$\text{MAE} = \frac{1}{n_T}\sum_{i=1}^{n_T}\lvert p_i - a_i\rvert$
Root mean square error	$\text{RMSE} = \sqrt{\frac{1}{n_T}\sum_{i=1}^{n_T}(p_i - a_i)^2}$
Relative absolute error	$\text{RAE} = \frac{\sum_{i=1}^{n_T}\lvert p_i - a_i\rvert}{\sum_{i=1}^{n_T}\lvert \overline{a}_i - a_i\rvert}$
Relative square error	$\text{RSE} = \frac{\sum_{i=1}^{n_T}(p_i - a_i)^2}{\sum_{i=1}^{n_T}(\overline{a}_i - a_i)^2}$
Co-efficient of determination	$\text{CoD}(R^2) = 1 - \sum_{i=1}^{n_T}\frac{(a_i - p_i)^2}{(a_i - \overline{a}_i)^2}$
Binary classification	
Precision (PRE)	$\text{PRE} = \frac{t_p}{t_p + f_p}$
Recall (REC)	$\text{REC} = \frac{t_p}{t_p + f_n}$
Accuracy (ACC)	$\text{ACC} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$
F1-score	$\text{F1} = 2 \cdot \frac{P_{RE} \cdot R_{EC}}{P_{RE} + R_{EC}}$
Area under RoC curve (AUC)	$\text{AUC} = \int_0^1 \text{RoC}$

Notations: n – total number of samples in the dataset, n_t – number of samples in the training dataset, n_t – number of samples in the test dataset, n_k – number of classes for binary classification, t_p – total number of true positive samples, t_n – total number of true negative samples, f_p – total number of false positive samples, f_n – total number of false negative samples, a –actual value, $\overline{a}$ – mean of actual values, p – predicted value, RoC – receiver operating characteristic curve.

4. Results and discussion

4.1. Data preprocessing

The raw dataset comprised of measured HR of 42 individuals and their corresponding MFCC frames. For each individual, we have a matrix of size 20 rows (coefficients in each frame) × 385 columns (frames) resulting in $n = 323,400$ coefficients for all the 42 individuals. It is understood that the HR of the individual remains unchanged during short time intervals, such as the duration of the speech segments in our dataset. Here, we have utilized the measured HR–MFCC dataset to develop a numeric HR prediction (regression) and a binary-classification model using machine learning statistical techniques.

Initially, feature ranking was performed to determine which MFCC frames are statistically significant for regression-classification study. We utilized the Filter Based Feature Selection (FBFS) module in MAMLS to score all the 385 MFCC frames in our dataset using Pearson's correlation coefficient score (Lin, 1989). Based on this score, it is found that more than 95% of the MFCC frames are statistically significant. Hence for our regression and classification study, we utilized all the 385 MFCC frames.

The dataset was also checked for any missing values in the extracted MFCC coefficients and was then normalized using MinMax normalizer to scale the MFCC coefficients in the range of [0,1] interval. Rows which had missing MFCC coefficients were discarded and not used in the analysis. Of the $42 \times 20 = 840$ rows of MFCC coefficients, 60 rows were discarded due to missing MFCC coefficients. This results in a total of 780 rows of MFCC coefficients with their corresponding HR values. A histogram of HR values corresponding to each of these MFCC coefficients is shown in Fig. 1. Normalization is achieved by shifting the values of each

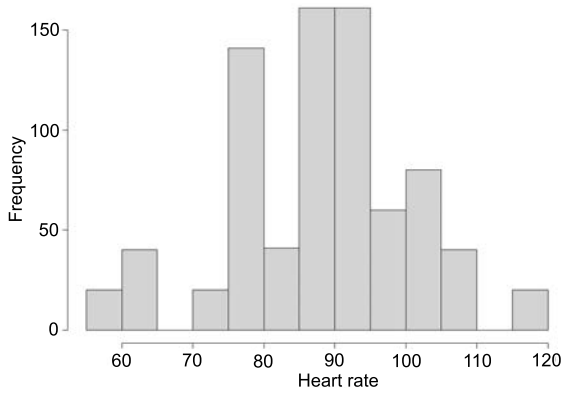

Fig. 1. Histogram of Heart Rate (HR) variable.

MFCC coefficient (denoted as x) so that the minimal value is 0, and dividing by the new maximal coefficient value, as follows:

$$\text{Normalized value} = \frac{x - \min(x)}{[\max(x) - \min(x)]}. \qquad (6)$$

Since the dataset has more MFCC frames than individuals samples, in this study, we have subjected the data to (n_T = 80%/n_t = 20%) split to ensure more samples are available for training and learning.

4.2. Regression analysis

The goal of this study is to apply regression machine learning algorithms mentioned in Subsec. 3.5, on the aforementioned dataset for predicting the HR of the individual from the MFCC coefficients extracted from speech signals. The schematic for the processing done post feature extraction is depicted in Fig. 2.

We trained two models using the optimally parameterized four regression algorithms, one without the predefined class data and the second with the predefined class data. It was observed that the HR prediction accuracy of the trained model with the inclusion of predefined class data is significantly higher for all the ML regression algorithms. The estimated heart rate obtained with and without HR class information along with the actual measured HR is shown in Figs 3–6 for

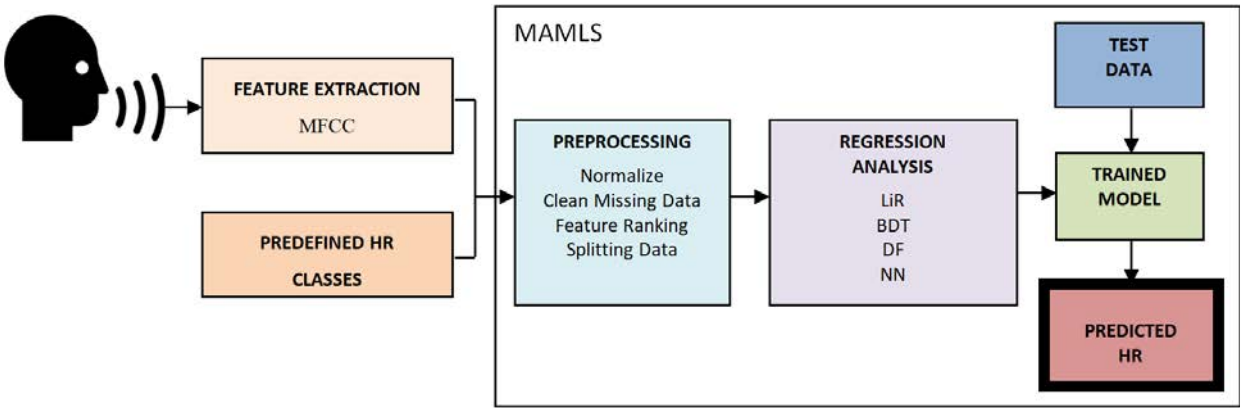

Fig. 2. Schematic representation of HR prediction.

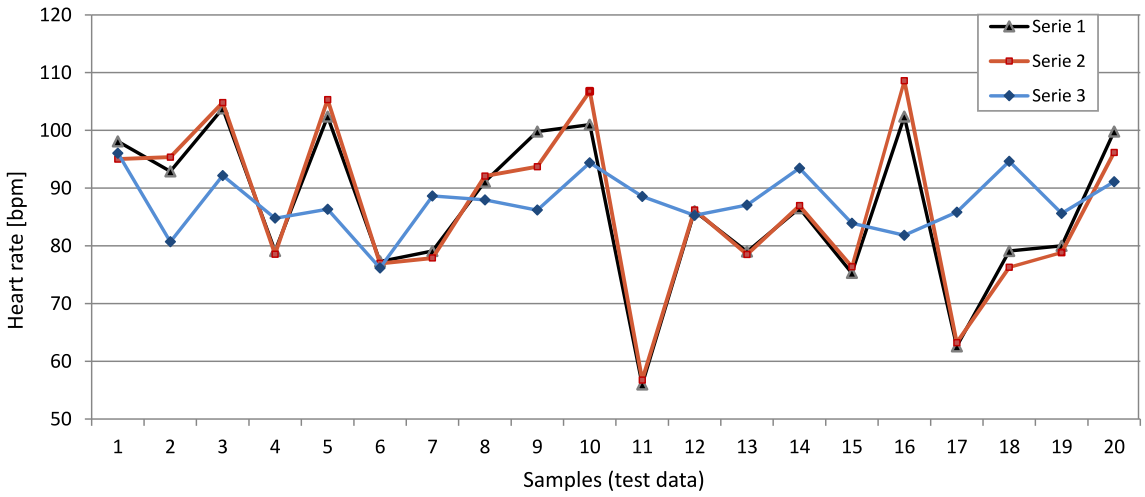

Fig. 3. HR regression using BDT algorithm.

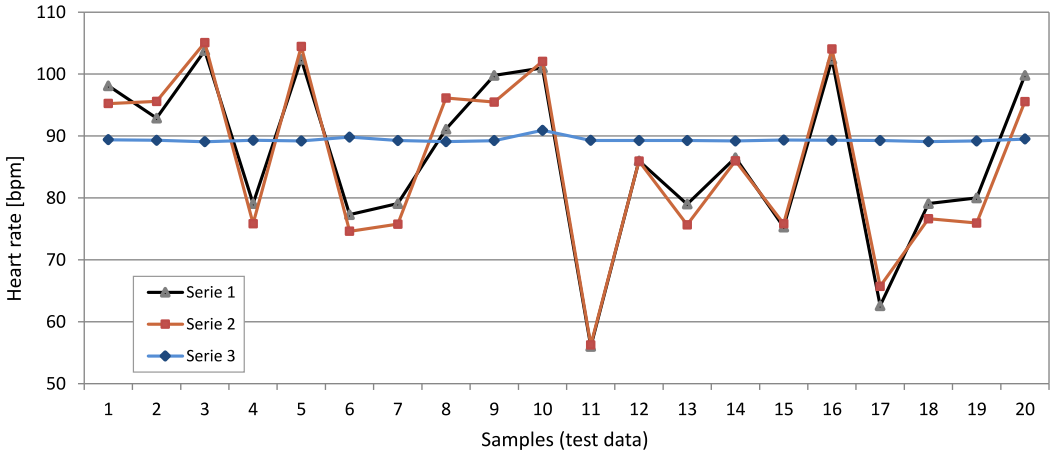

Fig. 4. HR regression using NN.

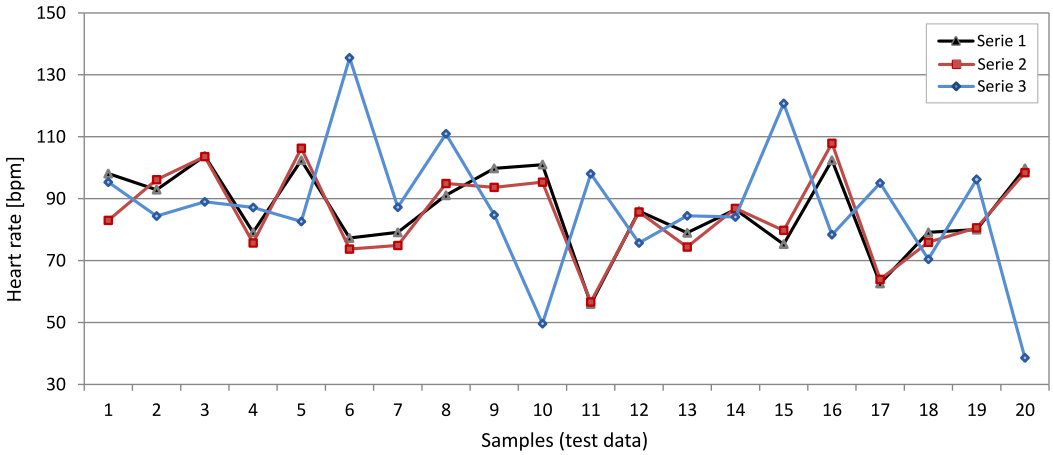

Fig. 5. HR regression using LiR algorithm.

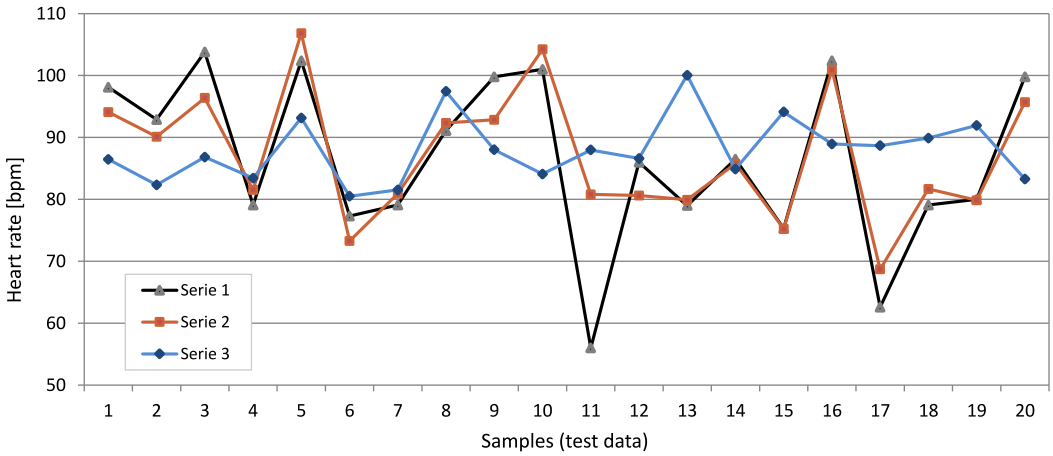

Fig. 6. HR regression using DF algorithm.

BDT, NN, LiR and DF algorithms respectively. The measured HR of the 42 individuals was divided into 5 classes as shown in Table 2.

All of the regression algorithms were optimally parameterized to achieve the best performance (the one with the lowest RMSE and highest R^2 values). For each of the aforementioned regression algorithms, the mean of the five evaluation metrics were computed after 10-fold cross validation, which are shown in Table 3, in which the standard deviation between folds for each of the performance metrics are listed inside round brackets. However, the coefficient of determina-

Table 2. Heart rate classes for regression analysis.

Pre-defined class	HR [bpm]
C0	50–60
C1	60–70
C2	70–80
C3	80–90
C4	90–100
C5	>100

Table 3. Performance metrics after 10-fold cross validation.

Regression algorithms	Optimal parameterization	Performance metrics Mean (standard deviation)				
		MAE	RMSE	RAE	RSE	CoD (R^2)
LiR	Solver – ordinary least square L2 regularization weight – 0.001	4.412 (0.452)	6.661 (1.283)	0.463 (0.049)	0.311 (0.141)	0.688 (0.141)
BDT	Number of leaves – 20 Learning rate – 0.09 No. of trees – 100	2.224 (0.064)	2.952 (0.123)	0.235 (0.030)	0.060 (0.013)	0.939 (0.013)
DF	Random split count – 128 Maximum depth – 32 No. of decision trees – 8	4.572 (0.399)	6.443 (0.761)	0.480 (0.038)	0.281 (0.046)	0.718 (0.046)
NN	Learning rate – 0.001 No. of hidden nodes – 257	3.542 (0.438)	4.661 (1.029)	0.373 (0.059)	0.159 (0.101)	0.840 (0.101)

tion (R^2) metric is widely used for exemplifying the predictive power of the regression model as a value between 0 and 1, with 1 being a perfect fit.

We plot in Fig. 7 the four regression models as a function of CoD (plotted as %) to predict the HR from the speech MFCC coefficient dataset. It is observed from Fig. 8 that the best performance (RMSE = 2.95, CoD = 0.94) is achieved for BDT algorithm. The 4 trained models were also compared on a test dataset consisting of 20 measured heart rate samples to predict the HR from the test data MFCC frames. A comparison of estimated HR values obtained used BDT, NN, LiR, and DF algorithms with actual measured HR values is shown in Fig. 8.

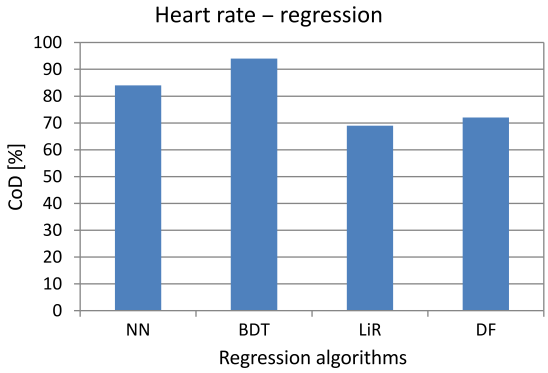

Fig. 7. Comparison of CoD [%] of different regression algorithms on HR estimation.

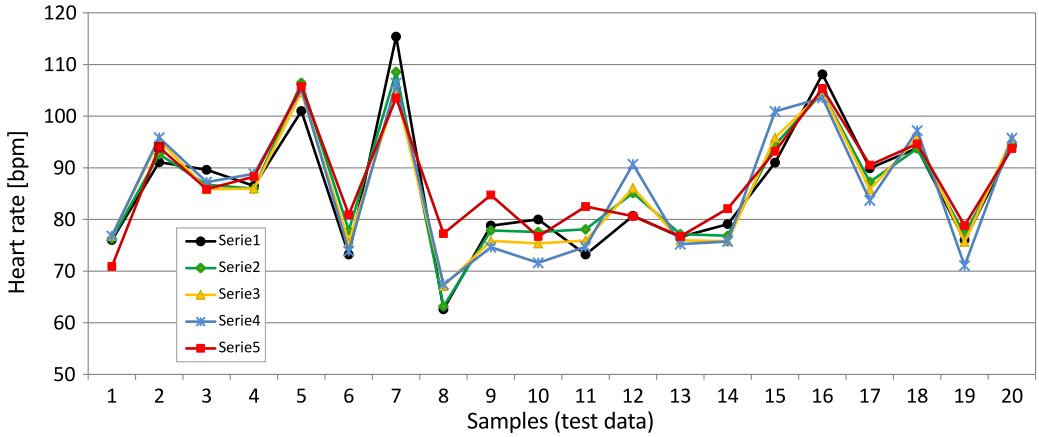

Fig. 8. Comparison of the measured (actual) and the predicted HR from the 4 trained models using BDT, NN, and DF regression algorithms on 20 test samples.

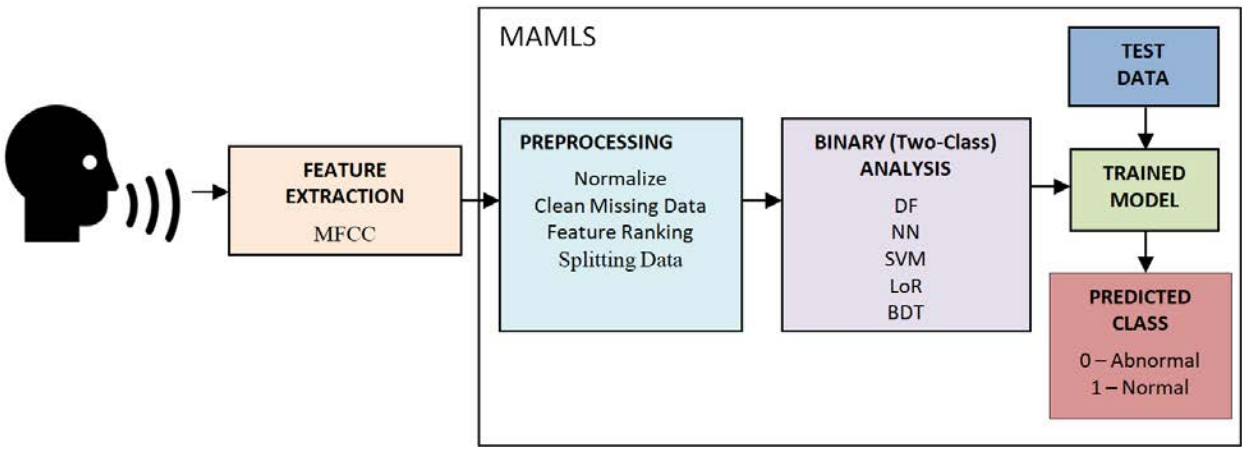

Fig. 9. Schematic representation of HR prediction.

4.3. Binary classification analysis

As shown in Fig. 9, for binary classification study, the measured resting HR of the 42 individuals in the dataset was divided into two binary classes namely: Class 1 for Normal (i.e. 60–100 bpm) and anything below 60 bpm and higher than 100 bpm was classed as Class 0 for Abnormal (Laskowski, 2018). The preprocessed HR-MFCC dataset utilized for regression was also applied to each of the aforementioned binary classification algorithms, we then compute the performance evaluation metrics i.e. P_{RE}, R_{EC}, A_{CC}, $F1$-score and the computed area under the curve (AUC) from the receiver operating characteristic curve (RoC) plot after 10-fold cross validation.

Optimal parameterization of each algorithm was performed to achieve the best classification performance (i.e. highest A_{CC}, $F1$-score) on the test dataset. The best accuracy (A_{CC} = 100%) and $F1$-score = 1 was again achieved using the BDT binary classification algorithm as observed in Table 4. Finally, we tested the models trained using the binary classification algorithms on a test dataset of 20 samples. It can be observed from Table 5 that the BDT trained model is able to accurately classify all the 20 test samples.

5. Conclusions

Speech signals contain intrinsic information regarding physiological, psychological as well as emotional conditions of the speaker. Accurate measurement of physiological parameters using speech signals can facilitate remote monitoring of patients and early diagnosis of medical conditions. The focus of this work is on estimating heart rate, which is a vital sign of individuals, from speech signals of the individuals. Heart rate estimation with high accuracy is achieved using speech spectral domain features (MFCC) as input to machine learning algorithms such as LiR, BDT, DF, and NN. HR estimation accuracy is highest for BDT algorithm. In addition to estimating the heart rate, a binary clas-

Table 4. Performance metrics computed for binary classification.

Classification algorithms	Optimal parameterization	Performance metrics Mean value (standard deviation)				
		A_{CC}	P_{RE}	R_{EC}	$F1$-score	AUC
DF	Random split count – 128 Maximum depth – 32 Number of decision trees – 8	0.891 (0.033)	0.898 (0.038)	0.973 (0.021)	0.934 (0.021)	0.92 (0.082)
NN	Learning rate – 0.001 Number of hidden nodes – 257	0.837 (0.057)	0.895 (0.049)	0.959 (0.025)	0.926 (0.02)	0.916 (0.079)
SVM	Lambda – 0.001	0.789 (0.072)	0.8091 (0.068)	0.962 (0.026)	0.877 (0.047)	0.812 (0.102)
LoR	Optimization tolerance – 1e–07 L1 regularization weight – 1 Memory size for L-BFGS – 20	0.839 (0.057)	0.831 (0.061)	1 (0)	0.907 (0.037)	0.881 (0.087)
BDT	Number of leaves – 20 Learning rate – 0.09 Number of trees – 100	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)

Table 5. Performance OF binary classification algorithms.

Test sample	Measured HR [bpm]	Actual class	Predicted class				
			BDT	DF	NN	LoR	SVM
1	82.9	1	1	1	1	1	1
2	93.8	1	1	1	1	1	1
3	78.8	1	1	1	1	1	1
4	91	1	1	1	1	1	1
5	76.6	1	1	1	1	1	1
6	108.1	0	0	1	1	1	1
7	78.8	1	1	1	1	1	1
8	115.4	0	0	0	1	0	1
9	115.4	0	0	0	1	0	1
10	88.9	1	1	1	1	1	1
11	82.1	1	1	1	1	1	1
12	91	1	1	1	1	1	1
13	89.9	1	1	1	1	1	1
14	76	1	1	1	1	1	1
15	103.8	0	0	0	1	1	1
16	89.6	1	1	1	1	1	1
17	78.8	1	1	1	1	1	1
18	94.8	1	1	1	1	1	1
19	91	1	1	1	1	1	1
20	76	1	1	1	1	1	1

sification scheme is also implemented to classify an individual's heart rate as 'normal' or 'abnormal'. Five techniques, BDT, DF, NN, LoR, and SVM, have been evaluated to address the classification problem. High accuracy is achieved for all the five techniques with DF having an accuracy close to 90% and BDT achieving 100% classification accuracy. Due to the unbalanced nature of the dataset used in this work, F1 score is a more indicative performance metric. Based on $F1$-score as well, BDT algorithm has the best classification performance followed by DF algorithm. Such high accuracies have been obtained by labeling the samples with predefined class information.

The proposed method has the following advantages over other methods available in the literature. In (Schuller *et al.*, 2013), classification accuracy of 82.7% and minimum MAE for HR estimation equal to 8.1 is reported. In comparison, the classification accuracy in this work is 100% using BDT algorithm and MAE is less than 5 for all the four algorithms used. While an accuracy greater than 95% is reported in (Mesleh *et al.*, 2012), it is restricted to only vowel sounds having a duration of at least 6 seconds and involves a lengthy procedure for each measurement. In contrast, the results in this work are not restricted to vowel sounds and once the AI algorithms are trained, the testing phase is relatively simple in terms of implementation complexity. It is indicated in (Kaur, Kaur, 2014), that accuracy of HR estimation from speech depends on various factors without actually quantifying it. Furthermore, voice recordings of 60 s duration are used as compared to less than 5 s segments in this work. The work in (Sakai, 2015b) is based on speech signals from only two individuals and the accuracy achieved is not specified. The classification of emotions based on speech MFCC features, reported in (James, 2015) exhibits large variation in accuracy across individuals. Compared to results available in literature, our results indicate better accuracy with fewer constraints. A limitation of this work is the small sample size (42) and lack of female speech samples which will be addressed going forward. It is intended to collect data from more individuals representing a much broader segment of the population which would further generalize the findings reported in this article.

Future work aims to achieve high accuracy without predefined class labeling and to detect atrial fibrillation for early detection of stroke. Measuring other physiological parameters such as blood pressure as well as monitoring of psychological and emotional conditions based on speech signals shall also be investigated in future. It is also intended to investigate the feasibility of using novel speech features, instead of MFCC's to measure physiological parameters from speech. The use of deep learning on raw speech signals rather than features extracted from speech signals shall also be in-

vestigated in future. The effect of varying the acoustic devices used for recording as well as varying the parameters of the recording devices is also a part of future work. Developing and training algorithms which are agnostic to the recording device will make the application of this work more useful and involves collecting data from individuals using multiple acoustic devices.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through General Research Project under grant number (G.R.P-352-39).

Authors thank Dr. Anis Ahmed of Blackpool Teaching Hospitals, NHS, United Kingdom, for his valuable inputs and suggestions.

References

1. Borkovec T.D, Wall R.L., Stone N.M. (1974), False Physiological Feedback and the Maintenance of Speech Anxiety, *Journal of Abnormal Psychology*, **83**(2): 164–168.
2. Bühlmann P., Yu B. (2003), Boosting with the L_2 loss, *Journal of the American Statistical Association*, **98**(462): 324–339, doi: 10.1198/016214503000125.
3. Burton D.A., Stokes K., Hall G.M. (2004), Physiological effects of exercise, *Continuing Education in Anaesthesia Critical Care & Pain*, **4**(6): 185–188, doi: 10.1093/bjaceaccp/mkh050.
4. Criminisi A., Shotton J., Konukoglu E. (2011), Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning, *Foundations and Trends® in Computer Graphics and Vision*, **7**(2–3): 81–227, doi: 10.1561/0600000035.
5. Davis S., Mermelstein P. (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**(4): 357–366, doi: 10.1109/ ASSP.1980.1163420.
6. Dreiseitl S., Ohno-Machado L. (2002), Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics*, **35**(5–6): 352–359, doi: 10.1016/S1532-0464(03)00034-0.
7. Euler C., Von (1982), Some aspects of speech breathing physiology, [in:] *Speech Motor Control. Proceedings of an International Symposium on Speech Motor Control*, Grillner S., Lindblom B., Lubker J., Persson A. [Eds], Stockholm, May 11–12, 1981, pp. 95–103, doi: 10.1016/B978-0-08-028892-5.50013-X.
8. Hermansky H., Morgan N. (1994), RASTA Processing of Speech, *IEEE Transactions on Speech and Audio Processing*, **2**(4): 578–589, doi: 10.1109/ 89.326616.
9. Hermansky H. (1990), Perceptual Linear Predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, **87**(4): 1738–1752, doi: 10.1121/1.399423.
10. Huang X., Acero A., Hon H.-W. (2001), *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR.
11. James A.P. (2015), Heart rate monitoring using human speech spectral features, *Human-Centric Computing and Information Sciences*, **5**(1): 1–12, doi: 10.1186/s13673-015-0052-z.
12. Kabal P. (2017), *Audio File Format Specifications*, MMSP Lab, McGill University, http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/CSL/CSL.html.
13. Kaur J., Kaur R. (2014), Extraction of heart rate parameters using speech analysis, *International Journal of Science and Research (IJSR)*, **3**(10): 1374–1376.
14. Kutner M.H., Nachtsheim C., Neter J., Li W. (2004), *Applied Linear Statistical Models*, 4th ed., Irwin: McGraw Hill.
15. Laskowski E.R. (2018), Heart Rate: What's Normal?, *Mayo Clinic*, https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979.
16. Lin L.I-K. (1989), A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, **45**(1): 255–268, doi: 10.2307/2532051.
17. Logan B. (2000), Mel frequency cepstral coefficients for music modeling, [in:] *1st International Symposium on Music Information Retrieval*, http://ismir2000.ismir.net/papers/logan_paper.pdf.
18. Lyons J. (2012), Mel Frequency Cepstral Coefficient (MFCC) Tutorial, *Practical Cryptography*, http://practicalcryptography.com/miscellaneous/machine-learning/ning/guide-mel-frequency-cepstral-coefficients-mfccs/#computing-the-mel-filter-bank.
19. MacGill M. (2017), Heart rate: what is a normal heart rate?, *Medical News Today*, https://www.medicalnewstoday.com/articles/235710.php.
20. Magre S., Deshmukh R.R., Shrishrimal P.P. (2013), A comparative study on feature extraction techniques in speech recognition, [in:] *International Conference on Recent Advances in Statistics and Their Applications*, Aurangabad, https://www.researchgate.net/publication/278549945_A_Comparative_Study_on_Feature_Extraction_Techniques_in_Speech_Recognition.
21. Merhav N., Lee C.-H. (1993), On the asymptotic statistical behavior of empirical cepstral coefficients, *IEEE Transactions on Signal Processing*, **41**(5): 1990–1993, doi: 10.1109/78.215323.
22. Mesleh A., Skopin D., Baglikov S., Quteishat A. (2012), Heart rate extraction from vowel speech signals, *Journal of Computer Science and Technology*, **27**(6): 1243–1251, doi: 10.1007/s11390-012-1300-6.

23. Nasrabadi N.M. (2007), Pattern recognition and machine learning, *Journal of Electronic Imaging*, **16**(4): 049901, doi: 10.1117/1.2819119.

24. Oppenheim A.V., Verghese G.C. (2015), *Signals, Systems & Inference*, Pearson.

25. Orlikoff R.F., Baken R.J. (1989), The effect of the heartbeat on vocal fundamental frequency perturbation, *Journal of Speech and Hearing Research*, **32**(3): 576–582, http://www.ncbi.nlm.nih.gov/pubmed/2779201.

26. Partila P., Voznak M., Mikulec M., Zdralek J. (2012), Fundamental frequency extraction method using central clipping and its importance for the classification of emotional state, *Advances in Electrical and Electronic Engineering*, **10**(4): 270–275, doi: 10.15598/aeee.v10i4.738.

27. Poh M.-Z., McDuff D.J., Picard R.W. (2011), Advancements in noncontact, multiparameter physiological measurements using a webcam, *IEEE Transactions on Biomedical Engineering*, **58**(1): 7–11, doi: 10.1109/TBME.2010.2086456.

28. Ramig L.A. (1983), Effects of physiological aging on vowel spectral noise, *Journal of Gerontology*, **38**(2): 223–225.

29. Reilly K.J, Moore C.A. (2003), Respiratory sinus arrhythmia during speech production, *Journal of Speech, Language, and Hearing Research: JSLHR*, **46**(1): 164–177, http://www.ncbi.nlm.nih.gov/pubmed/12647896.

30. Reynolds A., Paivio A. (1968), Cognitive and emotional determinants of speech, *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, **22**(3): 164–175.

31. Roychowdhury S., Bihis M. (2016), AG-MIC: Azure-based generalized flow for medical image classification, *IEEE Access*, **4**: 5243–5257, doi: 10.1109/ACCESS.2016.2605641.

32. Sakai M. (2015a), Feasibility study on blood pressure estimations from voice spectrum analysis, *International Journal of Computer Applications*, **109**(7): 39–43, doi: 10.5120/19204-0848.

33. Sakai M. (2015b), Modeling the relationship between heart rate and features of vocal frequency, *International Journal of Computer Applications*, **120**(6): 32–37, doi: 10.5120/21233-3986.

34. Schuller B., Friedmann F., Eyben F. (2013), 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*, pp. 7219–7223, doi: 10.1109/ICASSP.2013.6639064.

35. Schuller B., Friedmann F., Eyben F. (2014), 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-2014)*, pp. 1506–1510, Reykjavik: European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2014/pdf/611_Paper.pdf.

36. ScienceEncyclopedia (2019), Speech – the physiology of speech – air, vocal, words, and sound, *JRank Articles*, *Science Encyclopedia*, https://science.jrank.org/pages/6371/Speech-physiology-speech.html.

37. Scully C.G. *et al.* (2012), Physiological parameter monitoring from optical recordings with a mobile phone, *IEEE Transactions on Biomedical Engineering*, **59**(2): 303–306, doi: 10.1109/TBME.2011.2163157.

38. Skopin D.E., Baglikov S.U. (2009), Heartbeat feature extraction from vowel speech signal using 2D spectrum representation, [in:] *4th International Conference on Information Technology (ICIT)*, Amman, Jordan, https://www.zuj.edu.jo/conferences/ICIT09/PaperList/Papers/Image and Signal Processing/450Demitry.pdf.

39. Tan Z.-H., Lindberg B. (2010), Low-complexity variable frame rate analysis for speech recognition and voice activity detection, *IEEE Journal of Selected Topics in Signal Processing*, **4**(5): 798–807, doi: 10.1109/JSTSP.2010.2057192.

40. Trouvain J., Truong K.P. (2015), Prosodic characteristics of read speech before and after treadmill running, *16th Annual Conference of the International Speech Communication Association, Dresden, Germany* (ISCA), https://research.utwente.nl/en/publications/prosodic-characteristics-of-read-speech-before-and-after-treadmil.

41. Tufekci Z., Gowdy J.N. (2000), Feature extraction using discrete wavelet transform for speech recognition, [in:] *Proceedings of the IEEE SoutheastCon 200), "Preparing for The New Millennium"*, pp. 116–123, doi: 10.1109/SECON.2000.845444.

42. Usman M. (2017), On the performance degradation of speaker recognition system due to variation in speech characteristics caused by physiological changes, *International Journal of Computing and Digital Systems*, **6**(3): 119–127, doi: 10.12785/IJCDS/060303.

43. Usman M., Zubair M., Shiblee M., Rodrigues P., Jaffar S. (2018), Probabilistic modeling of speech in spectral domain using maximum likelihood estimation, *Symmetry*, **10**(12): 750, doi: 10.3390/sym10120750.

44. Wolf J.J. (1980), Speech signal processing and feature extraction, [in:] *Spoken Language Generation and Understanding*, pp. 103–128, Dordrecht: Springer Netherlands, doi: 10.1007/978-94-009-9091-3_6.

45. Yasuma F., Hayano J.-I. (2004), Respiratory sinus arrhythmia: why does the heartbeat synchronize with respiratory rhythm?, *Chest*, **125**(2): 683–690, http://www.ncbi.nlm.nih.gov/pubmed/14769752.

46. Zhang G., Patuwo B.E., Hu M.Y. (1998), Forecasting with artificial neural networks: the state of the art, *International Journal of Forecasting*, **14**(1): 35–62, doi: 10.1016/S0169-2070(97)00044-7.