FINGERTIP ELECTROCARDIOGRAM AND SPEECH SIGNAL BASED BIOMETRIC RECOGNITION SYSTEM

GÖKHAN GÜVEN

IŞIK UNIVERSITY
2021

FINGERTIP ELECTROCARDIOGRAM AND SPEECH SIGNAL BASED BIOMETRIC RECOGNITION SYSTEM

GÖKHAN GÜVEN
B.S., Electrical and Electronics Engineering, IŞIK UNIVERSITY, 2014
M.S., Electronics Engineering, IŞIK UNIVERSITY, 2016

Submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Electronics Engineering

IŞIK UNIVERSITY
2021

IŞIK UNIVERSITY
SCHOOL OF GRADUATE STUDIES

FINGERTIP ELECTROCARDIOGRAM AND SPEECH SIGNAL BASED BIOMETRIC RECOGNITION SYSTEM

GÖKHAN GÜVEN

APPROVED BY:

Prof. Dr. Ümit Güz (Thesis Supervisor)	Işık University
Prof. Dr. Hakan Gürkan (Thesis Co-advisor)	Bursa Technical University
Prof. Dr. Yorgo Istefanopulos	Işık University
Assoc. Prof. Cenk Demiroğlu	Özyeğin University
Assoc. Prof. Ramazan Köprü	Işık University
Assoc. Prof. Cemal Hanilçi	Bursa Technical University

APPROVAL DATE: 27../12/2021

FINGERTIP ELECTROCARDIOGRAM AND SPEECH SIGNAL BASED BIOMETRIC RECOGNITION SYSTEM

Abstract

In this research work, we presented a one-dimensional CNN-based person identification system which depends on the combination of both speech and ECG modalities to improve the overall performance compared to traditional systems. The proposed method has two approach: one is to develop combination of text-independent speech and fingertip ECG fusion system, the other one is to develop a robust rejection algorithm to prevent unauthorized access to the fusion system. In addition to the system robustness, we have developed an ECG spike and inconsistent beats removing algorithm, which detect and remove the problems caused by either portable fingertip ECG devices or movements of the patients.

First approach has been tested on 30, 45, 60, 75 and 90 people which were taken from LibriSpeech Corpus database and combination of both CYBHi and our private fingertip ECG database. The 3-fold cross validation test setup has been conducted while system working time was set to 10 seconds. In the first experiment, we achieved 90.22% accuracy rate for 90 people for ECG based system. For the speech based system, 97.94% accuracy rate has achieved for 90 people. For the combination of both system, 99.92% accuracy rate has been achieved.

For the second approach, 90 people for ECG and Speech database were being used as genuine class, 26 people as imposter class, and after the performance evaluation in optimum rejection thresholds, 71.08% accuracy rate for imposters rejection and 71.05% accuracy rate for genuine recognition has achieved for ECG based system. For the speech based system, imposter class were 87.82% accurately rejected while genuine classes were 86.48% accurately identified. The combination of both system has achieved 91.68% accuracy for genuine identification rate whereas 96.05% accuracy for imposter rejection.

Keywords: Authentication, Convolutional Neural Network, Fingertip ECG, Fusion, Identification, Imposter Rejection, MFCC, Machine Learning, Recognition System, Signal Processing, Speech, Supervised learning, Text-Independent, Verification

PARMAK UCU ELEKTROKARDİYOGRAM VE SES SİNYALİ TABANLI BİYOMETRİK TANIMA SİSTEMİ

Özet

Araştırmamızda, geleneksel sistemlere nazaran genel performansı iyileştirmek adına hem konuşma hem de EKG sinyallerinin kombinasyonuna dayanan tek boyutlu CNN tabanlı kişi tanıma sistemi geliştirilmiştir. Önerdiğimiz sistem, iki yaklaşım içermektedir: Bunlardan ilki, metinden bağımsız konuşma ve parmak ucu EKG füzyonu ile bir tanıma sistemi elde etmek, diğeri ise bu geliştirilen füzyon tanıma sisteminin yetkisiz kişileri önlemesine yarayan güçlü bir reddetme algoritması geliştirmektir. Bu yaklaşımlara ek olarak, taşınabilir parmak ucu EKG cihazlarının ya da kullanıcının hareketlerinin neden olduğu tutarsızlıkları veya benzeri sorunları tespit etmek ve ortadan kaldırmaya yarayan bir algoritma da geliştirilmiştir.

İlk yaklaşım, LibriSpeech Corpus ses veri tabanı ve CYBHi veri tabanı ile daha önceden oluşturduğumuz parmak ucu EKG veri tabanlarının birleşiminden alınan 30, 45, 60, 75 ve 90 kişi üzerinde test edilmiştir. 3 kat çapraz doğrulama yöntemiyle, sistem 10 saniyeye yanıt verecek şekilde ayarlanarak testler gerçekleştirilmiştir.İlk deneyde, EKG tabanlı sistemin, 90 kişi üzerinden %90.22 doğruluk oranına ulaştığı saptanmıştır. Konuşma tabanlı sistemin ise 90 kişi üzerinden %97.94 doğruluk oranına ulaştığı tespit edilmiştir. Her iki sinyalin kombinasyonu ise %99.92 doğruluk oranına sahip olduğu gözlemlenmiştir.

İkinci yaklaşımda ise, EKG ve konuşma veritabanlarından 90 kişi hakiki sınıf, 26 kişi ise sahtekar sınıfı olarak ikiye ayrılmıştır ve en uygun reddetme eşit değerlerine ayarlandığı göz önünde bulundurarak %71.05 doğrulukla hakiki sınıfı tanıdığı ve %71.08 doğrulukla sahtekar sınıfı reddettiği, EKG tabanlı sistemde tespit edilmiştir. Konuşma tabanlı sistemin ise, %86.48 doğrulukla hakiki sınıfı tanıdığı, %87.82 doğrulukla da sahtekar sınıfı reddettiği tespit edilmiştir. Her iki sistemin kombinasyonu ile, %91.68 doğrulukla hakiki sınıfı tanıdığı, %96.05 doğrulukla da sahtekar sınıfı reddettiği gözlemlenmiştir.

Anahtar kelimeler: Kimlik Doğrulama, Evrişimli Sinir Ağı, Parmak Ucu EKG, Füzyon, Tanımlama, Sahtekar Reddetme, MFCC, Makine Öğrenimi, Tanıma Sistemi, Sinyal İşleme, Konuşma, Denetimli öğrenme, Metinden Bağımsız, Doğrulama

Acknowledgements

There are many people who have helped to make my years at the graduate school most valuable. Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Ümit Güz, for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. Having the opportunity to work with him was intellectually rewarding and fulfilling. I also would like to thank Prof. Dr. Hakan Gürkan, my Ph.D. co-advisor, who contributed much to the development of this work.

Besides my supervisor and co-advisor, I would like to thank the rest of my thesis committee: Prof. Dr. Yorgo İstefanopulos, Assoc. Prof. Cenk Demiroğlu, Assoc. Prof. Ramazan Köprü and Assoc. Prof. Cemal Hanilçi for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

Last but not least, I would like to thank my family: my father Ziya Güven, my mother Ayla Güven, and my sister Göksu Güven for supporting me spiritually throughout writing this thesis and my life in general.

To my family and friends . . .

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1-D	**O**ne **D**imensional
2-D	**T**wo **D**imensional
AC	**A**uto**c**orrelation
AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
aVF	**A**ugmented **V**ector **F**oot
aVL	**A**ugmented **V**ector **L**eft
aVR	**A**ugmented **V**ector **R**ight
BPNN	**B**ack **P**ropagation **N**eural **N**etworks
CMRR	**C**ommon **M**ode **R**ejection **R**atio
CNN	**C**onvolutional **N**eural **N**etwork
CT	**C**ontinuous-**T**ime
DCT	**D**iscrete **C**osine **T**ransform
DMFCC	**D**ynamic **M**el-**F**requency **C**epstral **C**oefficient
DWT	**D**iscrete **W**avelet **T**ransform
ECG	**E**lectro**c**ardio**g**ram
EER	**E**qual **E**rror **R**ate

EMD	**E**mpirical **M**ode **D**ecomposition
EMI	**E**lectro**m**agnetic **I**nterference
ESD	**E**lectro**s**tatic **D**ischarge
FAR	**F**alse **A**cceptance **R**ate
FA	**F**alse **A**cceptance
FFBPNN	**F**eed-**f**orward **B**ack**p**ropagation **N**eural **N**etwork
FIR	**F**inite **I**mpulse **R**esponse
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FROS	**F**requency **R**ank **O**rder **S**tatistics
FRR	**F**alse **R**ejection **R**ate
FR	**F**alse **R**ejection
GBFS	**G**reedy **B**est **F**irst **S**earch
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel
HPE	**H**ermite **P**olynomials **E**xpansion
HRV	**H**eart **R**ate **V**ariability
IIR	**I**nfinite **I**mpulse **R**esponse
IMFCC	**I**nverse **M**el **F**requency **C**epstral **C**oefficients
KKCA	**K**ernel **C**anonical **C**orrelation **A**nalysis
KNN	**K**-**N**earest **N**eighbor

LDA	**L**inear **D**iscriminant **A**nalysis
LPC	**L**inear **P**redictive **C**oding
LRT	**L**ikelihood **R**atio **T**est
LSF	**L**ine **S**pectral **F**requencies
MEMs	**M**icro **E**lectro-**M**echanical **S**ystems
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MIT-BIH	**M**assachusetts **I**nstitute of **T**echnology-**B**eth **I**srael **H**ospital
MLP	**M**ulti-**l**ayer **P**erceptron
ML	**M**aximum **L**ikelihood
MOOC	**M**assive **O**pen **O**nline **C**ourse
MSE	**M**ean **S**quared **E**rror
PCA	**P**rincipal **C**omponent **A**nalysis
PTB	**P**hysikalisch-**T**echnische **B**undesanstalt
QRcp	**QR C**omposition with Column **P**ivoting
RBF	**R**adial **B**asis **F**unction
RELU	**Re**ctified **L**inear **U**nits
RFI	**R**adio **F**requency **I**nterference
RLS	**R**ecursive **L**east **S**quares
RNN	**R**ecurrent **N**eural **N**etwork
RPCANet	**R**obust **P**rinciple **C**omponent **A**nalysis **Net**work
SDMM	**S**uper-**D**irichlet **M**ixture **M**odel

SNR	**S**ignal to **N**oise **R**atio
SODP	**S**econd **O**rder **D**ifference **P**lot
STFT	**S**hort **T**ime **F**ourier **T**ransform
SVD	**S**ingular **V**alue **D**ecomposition
SVM	**S**upport **V**ector **M**achine
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
UBM	**U**niversal **B**ackground **M**odel
VAD	**V**oice **A**ctivity **D**etection
vMFMM	**V**on-**M**ises **F**isher **M**ixture **M**odel
VQ	**V**ector **Q**uantization

Chapter 1

Introduction

In thousands of years, body characteristics, such as face, voice, and gait, are being used by people to recognize each other. In the 19th century, Alphonse Bertillon, the chief of a criminal police department in Paris, developed a number of ideas about the measurement and identifications of criminals. Later on, his ideas were gained new popularities, and in the late 19th century, the distinctiveness of human fingerprints has been discovered. After the discovery of fingerprint, police departments began to store criminals' fingerprint in their database and relate them with the fingerprints which collected on the crime scene and identify the criminals. Although biometric recognition has emerged on extensive usage of criminal analysis, it is now increasingly used on a large number of civilian applications [1].

Any human physiological characteristics or behavior characteristic can be used as biometric parameters as long as it satisfies distinctiveness, universality, permanence, and collectability. However, there are several issues in practical biometric systems, such as performance which refers to speed and accuracy of the system; acceptability, which refers to the extent of people that are willing to accept the system; and circumvention, which refers to how easily the system can be deceived using fraudulent methods. Practical biometric systems should meet specific accuracy and response speed, be harmless to users, be accepted in a large population, and be sufficiently robust to different kinds of fraudulent methods [1].

Several biometric modalities have been accepted and are used in various applications. Each of the accepted biometric modalities has its strengths and weaknesses. The choice of modalities depends on the type of application. Description, advantages, and disadvantages of these modalities can be given as follows [1, 2]:

Iris: Iris is the annular region of the eye bounded by the pupil and the sclera (white of the eye) on either side. It is formed while fetal development and is stabilized after 2 years. The visual content of the iris texture has an advantage over other modalities in the potential of high accuracy, the resistance of imposter, long-term stability, and fast processing. However, iris can be easily replicated, can be used as an external, plus iris-based systems are primarily high-cost systems [1, 2].

Fingerprint: Fingerprint is the texture pattern over the surface of the fingertip, and it is formed and stabilized after 7 months of fetal development. The fingerprint-based system is easily used, has high accuracy, has long-term stability, the ability to enroll multiple fingers, and low cost compare to other modality-based systems. However, the system can be affected by the skin condition or dirtiness of the sensor's surface [1, 2].

Face: Facial-based systems are the most commonly used non-intrusive system. The most common approach is; finding the shape and location of the facial attributes such as eyes, eyebrows, nose, lips, or chin. Although this type of system is low-cost, the environment or appearances affect the system. It has a disadvantage of a highly false non-match rate, the potential of privacy abuse; identical twins attack [1, 2].

Signature: Signature is an individual's way of expression by signing his or her name on the paper. It is a widely accepted and non-intrusive method. This type of system has resistance over forgery. However, the system has problems over trivial and inconsistent signatures [1, 2].

Voice: Voice is the physiological and behavioral human biometrics. The human voice can change based on the shape and size of an individual's vocal tracts, mount, nasal cavities, and lips. This type of system can be easily applied by using existing telephony infrastructure or a simple microphone, has no negative effect on an individual. However, it can be easily fooled by pre-recorded individual voice, can be affected by background noise, or variability of voice when individual drunk or sick [1, 2].

Hand and Finger Geometry: Hand and finger geometry-based system view the features of the human hand, including its shape, size of a palm, length, and width of a finger, and palm lines. These type of system is easy to use, is not affected by environment, and it is relatively stable over fingerprint system. However, these systems have disadvantages of low accuracy, high cost, and difficulty to use for some users who have arthritis, missing fingers, or large hands [1, 2].

Electrocardiogram: Electrocardiogram (ECG) is a biological sign of the heart's electric activity and can be measured over human skins. It carries distinctive features depending on personal characteristics such as age, gender, size, position, and anatomy of the heart. This modality is being used not only because it is highly reliable, but it is also obligatory for an individual's presence. However, it is relatively hard to use if conventional devices are being used. Therefore, many kinds of acquisition systems were suggested and implemented to decrease the cost for the last 20 years. They focused on extracting ECG signals using an individual's hand or fingers rather than the chest to increase usability. However, this cost the accuracy of the system to decrease [2, 3, 4].

Many kinds of modalities are explained, and advantages, disadvantages between these modalities are mentioned. In recent years, many systems are suggested and developed to overcome these modalities' weaknesses by combining and fusing their strong aspect. For example, ECG based system has obligatory for an individual presence; therefore, it is hard to fool the system with pre-recorded devices. However, the performance of the ECG system decreases whenever the number of users

increases. Aging of the person or sudden emotional changes on people such as anger, fear, excitement can also affect the ECG-based system negatively. For this reason, the speech-based system used because speech does not drastically change over time and relatively stable over the ECG-based system. However, speech-based systems are easily fooled by using pre-recorded speech signals. Therefore, we proposed a system that focuses on both ECG and speech signal and developed a score fusion-based system by using their strong aspect.

In recent years, the popularity of ECG-based systems increases because of the high demand for security. In 2011, Z. Zhao and L. Yang [5] proposed an ECG-based algorithm that uses a matching pursuit algorithm to search the best time-frequency atom of each individual. After that, they performed the system using a Support Vector Machine classification algorithm to these individual features and achieved a 95.3% recognition rate over 20 subjects in the QT database. In the same year, Shen et al. [6] proposed a one-lead ECG human identification system and extracted ECG signals by using the palm of each subject. Then the database is constructed over 168 young college volunteers and used on an identification system that has a combination of template matching and pre-screening process. After extracting ECG feature templates, the Linear Discriminant Analysis classification method was performed on 168 people, and a 95.3% identification rate was achieved.

In 2012, Sara Zokaee et al. [7] developed and achieved a multimodal human identification system by using both palmprint and electrocardiogram. In their system, Mel-frequency Cepstral Coefficient (MFCC) method was being used to extract features for ECG signals, Principal Component Analysis (PCA) was being used to extract features for palmprint. Then for each of these features, distance similarities were found by using K-nearest Neighbors (KNN) classification algorithm and was tested on 50 people. Their system achieved a 100% recognition rate when only ECG signals were being used, which were taken from the MIT-BIH database, and was achieved an 89% recognition rate on the Dey-Hospital ECG database. Their system achieved an 82.1% recognition rate when palmprint was only being used,

and when both modalities are combined, their system achieved a 94.7% recognition rate. In the same year, Fufu Zeng et al. [8] proposed a statistical-based ECG identification method that applies the idea of matching Reduced Binary Pattern and was suggested of having high accuracy with low complexity and fast processing on ECG-based systems. P waves of each ECG signal were converted, reduced, and counted as binary sequences consisting of digits 0 and 1. After that, each of the ECG signal probabilities was being calculated by finding and giving them rank. The system was tested on MIT-BIH Normal Sinus Rhythm database consists of 13 women and 5 men, and was achieved a 90.19% identification rate. Their system achieved a 95.79% identification rate when MIT-BIH Arrhythmia Database consists of 25 men and 12 women were used.

In 2013, Emna Rabhi et al. [9] proposed a new set of features that took ten morphological descriptors of each heartbeat, which were divided into homogeneous groups such as amplitude, surface, slope, and intervals. These homogeneous groups were described as maximum amplitude of positive and negative peaks (Pp, Pn), area of positive and negative samples (ArP, ArN), the time interval from QRS complex to maximum positive peak, and time interval from QRS complex to maximum negative peak (Ima, Imi), QRS slope velocity between R peak to Q point, and R peak to S point (S1, S2). The system was constructed by combining these features with 60 Hermite Polynomials Expansion (HPE) coefficients which were extracted from each heartbeat. Then combination of these extracted features is applied on Hidden Markov Model. After that, the system was tested on 18 healthy people in the MIT-BIH database and achieved a 96.7% recognition rate. In the same year, A.C. Matos et al. [10] introduced an ARM cortex-based embedded ECG acquisition system. ECG signals of 10 subjects were measured by using dry Ag/AgCl electrodes on patients' left and right-hand index fingers. After that, ECG signals were segmented into 64 ms windows with an overlap of 54 ms, and Short Time Fourier Transform (STFT) was applied to each of these windows, where for each frequency bin, an estimation of mean and variance was stored in the database. The system compared the mean and variance

of each ECG signal with a pre-defined person's mean and variance by using the maximum likelihood estimation method. If the ECG signal was more likely to the pre-defined user for the given threshold, then it accepted the person; if not, it was rejected. 100% identification rate was achieved given 30 seconds ECG signal.

In 2014, J. Wu and Y. Zhang [11] presented a Neural Network-based ECG identification system which later tested on Lead-I ECG signals taken from 33 normal individuals in the MIT-BIH Arrhythmia Database. QRS complexes were suggested that is the most distinct feature sets among the heartbeat features and were being used for human identification. Their system consists of extracting QRS complexes on each heartbeat, then applying the dimension reduction method by using the PCA algorithm on the QRS complexes. Reduced QRS complexes were then applied on Back Propagation Neural Networks (BPNN) as a classifier to score and evaluate the system's performance. In the research, the accuracy of the classifier reached 99.6% for 33 subjects. In the same year, Kuo-Kun Tseng et al.[12] proposed an identification system by constructing a sparse matrix that contained ECG signals that reduced dimensions. Through their survey, transform-based such as wavelet transform frequency domain transform, discrete cosine transform (DCT), and waveform-based feature extraction methods were being observed, and the best feature extraction method was decided as a waveform-based method for their system. The unique aspect of their research was, not only Lead I ECG signals were being used, but also other leads were being considered as features because of containing additional information of the heart. This idea formed from research [13] focused on diagnosing acute myocardial infarction using two-dimensional echocardiography. In their research, two lead ECG signals were being mapped into two-dimensional coordinates to form a matrix. After that, a special mask was applied onto the matrix, which desired to be reduced in dimension so that it can transfer into a sparse matrix that easily stored and addressed the signals. The important point of the sparse matrix was that it reduced and stored the non-zero elements into three vectors which contained rows of the non-zero elements, the column of the non-zero elements, and

values of the non-zero elements. After finding the rows, columns, and values of non-zero elements, coordinate format (COO) was being used to store the system; as most of the matrix were zero, the corresponding value '1' was given. Then correlation coefficient classifier was being used for the purpose of training and individual identification. In the training process, for each of ECG sparse matrix's correlation and covariance were found among each other, and the best threshold values were being found for each person so that in the test process, it was used to classify the unknown ECG data. Accuracy of 95.3%, False Acceptance (FA) of 0.094, False Rejection (FR) of 0 was achieved by using 18 long-term ECG recordings in the MIT-BIH normal sinus rhythm database.

In 2015, Huan Zhang et al. [14] proposed an ECG Identification system that fused two different feature extraction methods, which were harmonic fundamental wave ratio and single-cycle inner time-frequency joint analysis. In the pre-processing, a wavelet de-noising algorithm was applied on the ECG signal of 36 individuals in the MIT-BIH database so that muscular noise on ECG signals was reduced. After the noise reduction algorithm, R points of ECG signals were found, and two adjacent R points were defined as one cardiac cycle. Firstly, a total of 32 harmonics were found by using discrete Fourier Series expansion for each cardiac cycle. The taking mean of every 32 harmonics found in each consecutive ten cardiac cycles was then stored as a feature vector to achieve minimum diversity. In their experiment, 32 harmonics were the best value that fits the ECG waveform well. Then Short Time Fourier Transform (STFT) was applied between each QRS complex, and new feature sets were found. After that, by combining these two feature sets, a 94.4% recognition rate was achieved. In the same year, Juan Sebastian et al.[15] introduced an ECG authentication system for mobile devices. The idea comes from the excessive need for security when compares with traditional password systems that are not enough for protection. With this in mind, they suggested an authentication system that users press two metal electrodes externally attached behind the cellphones to unlock their cellphones. In the algorithm, fiducial points were extracted from ECG signals, which were P point,

Q point, R point, S point, T point, the starting point of the P wave (LP), and ending point of T wave (TP). The idea comes from research [16] focuses on finding the best temporal fiducial points for ECG identification system. In the laboratory experiment, the algorithm was tested on ten different people that recorded at different times and achieved a 1.41% false acceptance rate and 81.8% true acceptance rate for 4 second recorded signals. In the same year, M. Dai et al. [17] proposed a weighted correlation coefficient method for ECG-based identification. Their research stated that correlation coefficient was the statistical indicator that reflected the linear relation between two variables; by calculation it, determination and classification of ECG waveforms of the person can be achieved. In their research, the contribution of the correlation coefficient was under debate given the ECG template. Therefore they introduced a weighted correlation coefficient system and achieved a recognition rate of 98% compare to traditional correlation coefficient methods, which have 94.15% in the MIT-BIT database. It achieved an 87.67% recognition rate, where the traditional method achieved 77.15% when ECG signals were recorded from hands. In the same year, G. Altan et al. [18] proposed an identification system that uses the Second-order Difference Plotting (SODP) method to extract features for each segment of the heartbeat. It is mentioned in their research that SODP is a method that analyses non-linear signals by visualizing two consecutive data points. Each segment found from ECG was mapped into two dimensional coordinates by considering x(n+1)-x(n) for Y-axis, x(n+2)-x(n+1) for X-axis. Then classification system was constructed by using the K-nearest Neighbor algorithm. After that, their system was performed on 90 people's ECG signals which were taken from the Physionet database and achieved a 91.52% recognition rate. In the same year, Muhammad Najam Dar et al. [19] presented an ECG recognition system based on hybrid feature extraction methods, which are heart rate variability (HRV) and discrete wavelet transform (DWT). First local maxima of ECG signals were found to extract R peaks, then 45 samples left, and 49 samples right side of the R peaks were defined as QRS complex. After locating the QRS complex, discrete wavelet (Haar) transform was applied on them so that extract more discriminating features for each beat.

Then corresponding heart rates, which are time differences between R-R peak interval and mean of R-R interval, were extracted as for time-domain features. Greedy Best First Search (GBFS) was used to determine highly correlated feature sets among the combined time and frequency domain features. Then it achieved 95.85% accuracy, 4.15% false acceptance rate, and 0.1% false rejection rate on 47 subjects by using Random Forest classifier.

In 2016, Xiafei Lei et al. [20] represented a deep learning feature-based system that reduced the dependence of origin and length of the ECG signals on system accuracy. Unlike traditional methods, feature combinations and fiducial point detection were not required in their system, and it was stated that the test process could use parallel processing to improve overall efficiency. Firstly, a 1 Hz – 40 Hz bandpass filter was applied to ECG signals to remove unwanted noise. After that sampling rate of ECG signals was decreased to 125 Hz to increase deep neural network speed. Then a window with a length of d-window was pre-defined, and ECG signals were segmented into vectors by the size of d-windows. Their system consists of 1 dimensional two different convolutional neural networks. First CNN was applied to analyze the temporal points for the given segmented ECG vectors, and it had 1 input layer with the dimension of d-window, three convolutional layers with 5 x 1, 5 x 1, and 4 x 1 filter lengths, respectively, average pooling layer with a 2 x 1 subsampling step, one fully connected which used as feature 1. It was stated that convolutional layers are equipped with randomized leaky rectified linear units (RReLU). After that, in the parallel process, the second CNN, which had the same structure as the previous one, was applied to the coefficients found by applying Discrete Cosine Transform (DCT) on segmented ECG vectors and was used as feature 2. After combining these two features, Back Propagation Neural Network (BPNN) and a non-linear Support Vector Machine (SVM) were used as classifiers. BPNN had 3 layers with 200 neurons in the hidden layers, whereas the Gaussian radial basis function (RBF) kernel was being used on SVM. The system was tested on 100 subjects who were randomly selected from the PTB Diagnostic ECG Database and achieved 99.33% accuracy. In the

same year, Kuo-Kun Tseng et al. [21] proposed a Neural Network-based unknown ECG identification system in which Frequency Rank Order Statistics (FROS) as a feature extraction method and BPNN as classifier were used. In FROS, firstly, ECG signals were mapped to binary number by comparing with the adjacent sample. '0' was given when the value of the adjacent sample was bigger than the current sample, whereas '1' was given when the value of the adjacent sample was smaller than the current sample. After the ECG signals were converted to binary numbers, they were converted to m-bit words by the length of m, which will be counted and sorted in order of descending frequency. Then ECG ranks were given according to their frequency, from the largest to smallest, and were used as feature sets. System performance was tested on 18 subjects taken from Arrhythmia Laboratory at Boston's Beth Israel Hospital. Then 1 person was used as an unknown class, the other 17 people were used for training the Neural Network, and an unknown class was identified with an accuracy of 91.76%. In the same year, M. Bassiouni et al. [22] presented a person identification method by using an artificial neural network (ANN) on non-fiducial ECG features. ECG signals were separated into vectors by the 10 seconds window length. After that, autocorrelation (AC) of each of these separated ECG signals were calculated, then discrete cosine transform (DCT) was applied on the AC and found AC/DCT features of each separated ECG signal. These AC/DCT coefficients are then used on ANN to construct a classification algorithm. 30 subjects in the MIT-BIH Arrhythmia database were being used for performance, and accuracy of 97% was achieved.

In 2017, L. Wieclaw et al. [23] proposed a biometric identification system based on deep learning techniques. First, the raw ECG database was constructed from the subject's three fingers by using Ag/AgCl electrodes to make the system more user-friendly. Then various digital filters were applied on raw ECG signals so that baseline wanders, 50 Hz sinusoidal interference, muscle artifacts including respiration movement, and noise generated from the electronic device would be eliminated. Then each of the ECG signal's R points was found, segmented into

a fixed dimension, and used on a feed-forward network to train the system. The system performed on raw ECG signals of 18 people and achieved a 96% identification rate. In the same year, Gang Zheng et al. [24] also proposed a biometric identification system based on deep learning techniques. The difference is, the feature set was found from information entropy. This information entropy is the measurement of confusion and dispersion degree of ECG signals of same people to others. Their system was performed on three different databases, which consisted of 13 individuals from the MITDB database, 28 individuals from self-collected, 19 individuals with different emotion statuses. Respectively, 96.63%, 98.10%, and approximately 95.51% recognition rate were achieved. In the same year, Ronald Salloum et al. [25] also presented an ECG identification method in which recurrent neural network (RNN) was used. Firstly, their system was found each heartbeat for a person and put them into fixed N x D vectors where N represented the number of heartbeats, whereas 'D' represented the dimension of each heartbeat waveform. After that, newly constructed 2-D features were given to the RNN algorithm for classification. Their system was tested on different dimensions, and 9 heartbeats were selected because of giving the best performance. Then it was performed on two different databases, which were ECG-ID and MIT-BIH Arrhythmia databases that; the number of classes consisted of 18 individuals and was achieved approximately 0% EER and 100% recognition rate when 9 heartbeats were being used for each individual.

In 2018, M. Bassiouni et al. [26] introduced intelligent hybrid approaches for human ECG signal identification. Firstly, the de-noising and filtering process was applied on ECG signals which were taken from the MIT-BIH Arrhythmia database and ECG-ID database. Then both fiducial features, which were P, R, T waves amplitude, Q and S point amplitude, and their respective intervals (QRS, P-QRS, QRS-T, P-QRS-T) and non-fiducial features were extracted. AC/DCT method, which was explained in their previous work, was applied for finding non-fiducial features. After the combination of these two approaches, three different classification methods were being used, which were SVM, ANN, KNN, and 99%,

95%, 98% identification rates were achieved on 90 individuals in the ECG-ID database, respectively. For 47 people in the MIT-BIH database, each method achieved a 100% identification rate.

In 2019, Jae-Neung Lee et al. [27] presented a personal identification system that used a robust principal component analysis network (RPCANet) and wavelet analysis. First, the wavelet transform of ECG signals was found by taking the inner product between the original signal and the so-called wavelet basis function. Newly found wavelet coefficients gave us similarities between analyzing and analyzed signal, and these coefficients indicate how close the signal was to the basis function. Scalogram was applied to wavelet coefficient to visualize it and mapped it into two dimensions. It allowed us to detect the most representative frequencies and contribution the most to the total energy. After that, RPCANet was applied to the 2-D features by convolving them with PCANet filter banks and by using multiple binary quantization operations to scale it to 1-D sparse features. SVM was used on these 1-D sparse features, and a 98.25% recognition rate was achieved on 95 individuals in their self-constructed database.

These are the most recent research about ECG identification. Before we begin the speech-based identification, some concepts must be known. The speech-based identification method is divided into two different categories, which are text-dependent speaker recognition and text-independent speaker recognition. Text-dependent speaker verification is connected with the fact that a pre-defined utterance is used for both training and testing the system. It can also be stated as Fixed Phrase Verification, in which a pre-defined phrase is used both during the training and verification periods. In comparison, text-independent speaker identification is not restricted to any fixed and prompted phrases [28]. Therefore, although this type of system has more freedom on users, the system's performance is not that accurate with respect to a text-dependent system. Text-independent methods mostly focused on spectral characteristics of each speaker by extracting one or more codebook entities that were presentative of that speaker. We will discuss with text-independent system and its recent development because

the text-independent speaker identification method is being used in our proposed method.

In 2007, Sandipan Chakroborty et al. [29] proposed a text-independent speaker identification method by combining MFCC with a newly developed filter bank structure. Over the years, MFCC was being used to model the human auditory system. However, because of the structure of the standard MFCC filter bank, it captured vocal tract characteristics in the lower frequencies. In their system, a new set of features were extracted by using an additional complimentary filter bank structure which improved the distinguishability of speakers, specific in higher frequency zone. These newly found MFCC coefficients were stated as Inverse MFCC (IMFCC) because the filter bank was the inverse of the standard filter bank. Additional to IMFCC, MFCC was also found, and these two features were then put into Gaussian Mixture Model (GMM) classifier. GMM was a non-parametric classifier, and it was defined as a multivariate probability distribution model which was capable of modeling the arbitrary distribution of speaker. During the training process, MFCC and IMFCC features were given to the Expectation and Maximization algorithm, which iteratively updates the parameters until the log-likelihood converges to a stable value. Their system was tested on 138 speakers, each has 96 utterances in the YOHO database, and a 97.74% recognition rate was achieved. In the Polycost database, their system was achieved an 81.57% recognition rate.

In 2009, Shung-Yung Lung [30] proposed a text-independent speaker recognition system based on improved wavelet feature extraction using kernel analysis. In general, discrete wavelet coefficients are not diagonalizable with respect to wavelet bases, and these schemes may lead to an eigenvalues decomposition of a very large covariance matrix which is computationally expensive. For this reason, the kernel canonical correlation analysis (KKCA) was being proposed to conduct after the wavelet transform. After the feature was extracted, GMM was being used as a classifier. 100 telephone data in the TALUNG database, which consists of 65 male, 35 female speakers who pronounced free-text, have been used and achieved a

96% recognition rate. In the King database, which consisted of 51 male speakers, a 98% recognition rate was achieved.

In 2010, Sandipan Chakroborty et al. [31] introduced their new methods about text-independent speech recognition by stating that eliminating redundant features was important for the performance. They stated that redundant features confuse the speaker model in multidimensional space resulting in degrading performance. Carefully selecting features is not only helping achieved a higher rate of accuracy but also lowers the computational cost. For this purpose, Singular Value Decomposition (SVD) followed by QR Composition with Column Pivoting (QRcp) methods applied to the features. Three different feature extraction methods had been used, which were MFCC, Linear Frequency Cepstral Coefficients (LFCC), and a newly proposed method with a Gaussian shape filter on Mel-scale (GMFCC) for comparing each other. The procedure for finding GMFCC was exactly the same as MFCC expect for using triangle filter bank on Mel-scale, Gaussian shape filter bank was being used. It was stated that GMFCC did not only contain frequency information like MFCC, it was also carrying amplitude information derived from the power spectrum. GMM was used as a classifier and, the system tested on 131 people in Polycost and 138 people in the Yolo database. By using MFCC, LFCC, and GMFCC features, their system achieved 96.65%, 96,75% and 97.03% recognition rates on the Polycost database, respectively. However, In the Yolo database, their system achieved 77.85%, 77.85%, and 80.24% recognition rates, respectively. In the same year, M. S. Sinith et al. [32] proposed text-independent speaker identification using MFCC and GMM. Their system tested on 60 speech data which consisted of a combination of English, Hindi, Tamil, and Malayalam languages, and achieved 98.8% recognition rate when tested speech signal lengths were 10 seconds.

In 2011, Zhanyu Ma et al. [33] represented a super-Dirichlet mixture model using differential line spectral frequencies for text-independent speaker identification. Their system utilized the line spectral frequencies (LSFs) as an alternative feature

set for capturing speaker characteristics. Then LSF was transformed to the differential (DLSF) space by cascading two neighboring frames which one of which was past frames, whereas the other one was following and current frames. Then combination of three vectors gave us a super vector, and the statistical distribution of this super vector was modeled by the super-Dirichlet mixture model (SDMM). The feature extraction procedure was exactly the same as MFCC until three consecutive MFCC features were put to 3 x M vector where M was the dimension of MFCC feature set while applying linear regression for each MFCCs. Then super vector was mapped to a 1 x N vector by using the DLSF method. It was stated that the proposed model had been achieved promising improvements. Their system tested on 25 people in the TIMIT database where each speaker spoke ten sentences, and in the pre-processing, speech data was segmented into 25 ms frames which have 10 ms step size, and silence frames were removed. It achieved a 99.5% recognition rate by using an SDMM classifier. In the same year, Khaled Daqrouq [34] proposed a text-independent speaker recognition system by using wavelet entropy and neural network. By using Shannon entropy of wavelet packet (WP), 4 feature sets were extracted from speech. Then, the features fed to the feed-forward backpropagation neural network (FFBPNN) to classify the speech signals. The system tested on 29 speakers who utter '0' to '14' digits and a total of 696 utterances in the Arabic language. The first 6 utterances were being used for the training set, while others were being used for testing the system for each person. In the experiment, a 91.09% recognition rate was achieved. In the same year, Hesham Tolda [35] proposed a text-independent speaker identification method based on MFCC and Continuous-Time Hidden Markov Model (CT-HMM) with Gaussian Mixture Model. It stated that Hidden Markov Model was a non-parametric model, where the set of short-time training feature vectors of a speaker compressed to a small set of representative points. CT-HMM was an HMM in which both the transitions between hidden states and the arrival of observations could occur at arbitrary times. Therefore, it is suitable for irregularly sampled temporal data such as speech signals. However, it comes with a

more computational cost. The system tested on 10 speakers whose speech consists of two sentences recorded in different rooms. The first sentence was being used on training, and the other one that collected from a different room used to test the system, and approximately 80% recognition rate was achieved whereas same sentences were used on both test and training, it was stated that it achieved approximately 100% recognition rate.

In 2012, A.D. Jafeer et al. [36] proposed text-independent speech recognition using MFCC and Hidden Markov Model. MFCC features were first put into vector quantization so that redundant features could be eliminated. Then reduced features were pass through HMM for classification purposes. 40 speakers with 20 utterances per speaker from the Switchboard Corpus dataset were being used, and each utterance consisted of approximately 3.2 seconds speech signal. The average length of each frame was about 32 ms, so each speaker had at least 3960 MFCC feature vectors. In both test and training datasets, 128 significant MFCC features were extracted from a total of 3960 MFCC feature set for each person by using the K-mean clustering algorithm. Three different classifiers, which were Linear Discriminant Analysis (LDA), Multilayer Perceptron, and Hidden Markov Model, were being used, and 70.5%, 81.1%, and 100% recognition rate were achieved, respectively.

In 2013, S.S. Nidhyananthan et al. [37] proposed robust language and text-independent speaker identification, which used a combination of Dynamic Mel-frequency Cepstral Coefficient (DMFCC) feature and MFCC feature. It was stated that pitch frequency presented the speaker's periodic characteristic of the vocal cord's vibration when speakers pronounced voice sounds, and because the traditional MFCC algorithm used a fixed-size filter bank, it did not fully portray the vocal characteristics of different people. Therefore dynamic-Mel filter-based feature extraction algorithm was represented. Their system tested on 120 people who have a speech length of 20 seconds. By using GMM classification, their system achieved 1.2% error rates when the combination of DMFCC and MFCC was used, whereas 5.8% and 2.9% error rates were achieved when only MFCC

or DMFCC were used respectively. In the same year, Jalil Taghia et al. [38] presented Von-Mises Fisher Mixture Model (vMFMM) based text-independent speaker identification method. LSF algorithm was used for feature extraction, and Von-Misses Fisher Mixture Model was used for classification. If the vectors in which 'Direction' is more important than 'Magnitude' of the datasets, then this classifier is being used. It models the signal based on distribution on a hypersphere and mainly consists of two parameters which are mean direction (μ) and concentration (κ). The algorithm starts with the initialization of the hyper-parameters characterizing the parameter distributions. Then, the current distribution over the model parameters is being used to evaluate the responsibilities, which will be used for the optimization of the variational posterior distribution over parameters. Variational lower bound will be monitored for each iteration until converging to a specific value. Finally, the predictive density of the new observed variable is calculated and used for classifying the system. TIMIT speech database was used for the evaluation of the system. 100 randomly selected speech signals were segmented into 25 ms frames with a length of 10 ms step size. Then silence frames were removed from the database. It achieved a 78.43% recognition rate on 3 seconds speech signals when vMFMM and LFSs were being used, whereas a 77.36% recognition rate was achieved when GMM and MFCC were being used.

In 2014, Hong Yu et al. [39] proposed a text-independent speaker identification method that used a histogram transform model on MFCC features. Their system used dynamic MFCC features, which took and put adjacent frames into a super vector and calculated 3 neighboring MFCCs. Then probability density function (PDF) of the super vector was estimated by using the histogram transform (HT) algorithm. In the test procedure, seven sentences that were randomly selected from one speaker were used as training, and the remaining three sentences for each speaker were used for testing. When the super vector contained consecutive 100 MFCCs, it achieved a recognition rate of 97.6%. In the same year, N. Almaadeed et al. [40] proposed a text-independent speaker identification system based on wavelet analysis and neural network. Their system consisted of three

neural networks, which were RBF, PNN, and GRNN, and score fusion was established by concerning majority vote. The system tested on 34 people in the Grid database, which had 1000 sentences spoken. The signals were separated into 50 ms frames and applied to classification methods, and a 97.5% identification rate was achieved.

In 2016, N. M. AboElenein et al. [41] introduced improved text-independent algorithm by using MFCC and GMM methods. CHAINS Speech Corpus dataset, which consisted of 36 people who were recorded with a time separation of 2 months, had been used in their system. Each signal passed through pre-process stage, and in the pre-processing, downsampling and silence removing operations were conducted. After that gender detection algorithm was used for separating females and males to improve accuracy; for that reason, a pitch detection algorithm was used, voiced and unvoiced parts of speech signal were found. Then by using the voiced part of the speech signal, gender of speech was detected. After that, MFCC features were extracted for both male and female speech, separately. Then vector quantization was applied for removing feature redundancy and model with GMM algorithm. Their proposed system was achieved a 91% recognition rate, whereas the traditional system, which uses only VQ and GMM methods, achieved an 88% recognition rate. It was stated that the time it consumes was less than 20% from the traditional system. In the same year, N. Almaadeed et al. [42] proposed a vowel-based recognition algorithm for real-time text-independent speaker identification. The system contributed to designing a scalable system based on vowel formants filters and a scoring scheme for the classification of an unseen instance. Both MFCC and Linear Predictive Algorithm (LPC) had been used to extract vowel formants in given 30 ms speech frames. LPC is a method that extracts the resonating frequencies of formants from the remainder of the noisy signal through inverse filtering. It analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating their intensity and frequency. It was stated that frequency range and standard deviation of vowel formants were found by the research of L.R. Rabiner et al. [43], and by

series of bandpass filters, vowel formants can be found. In their system, the Window process, Normalization, Auto-Regression filter, Formant detection method, Matched Filter/Smoother algorithms were applied in order to find vowel formants. 5 different features were extracted and used as score-based speaker identification. These 5 different features were given as both first and second formants, all first, second, third formants, both first and second formants, both second, third formants, and averaging, comparison with least distance. These 5 feature sets were then given to neural networks to score based identification. The system tested on four databases which are YOHO, NIST, TI_digits1, and TI_digits2. It was only stated TI_digits databases consisted of 326 speakers who utter 77 digit sequences. All speech files, each at least 25 seconds long, had been used for training the system, and it was stated that above the 25 seconds speech data was hardly make any difference to the performance of the system. The system was tested on 1, 2, and 3 seconds speech signals, and for the YOHO database, it achieved 83.78%, 89.46%, 94.23% recognition rate, respectively. For NIST database, it achieved 72.54%, 85.28%, 92.15% recognition rate, respectively. For TI_digits1 database, it achieved 78.56%, 91.25%, 96.87% recognition rate, respectively. And finally for TI_digits2 database, it achieved 80.12%, 91.89%, 97.34% recognition rate, respectively.

For clarifying the differences between the voiced, unvoiced, and vowel formants, A.G. Ramakrishnan in Indian Institute of Science stated that "Pitch is the fundamental frequency of vibration of the vocal folds, which are present at the top of one's trachea. They vibrate quasi-periodically only for voiced phonemes, namely vowel, semivowel, and nasal sounds. So, for unvoiced stops such as /p/, /k/, /t/, /th/, /ch/ and unvoiced fricatives such as /f/, /s/, etc. there is nothing called pitch. The formant frequencies are due to the frequency shaping of the signal from the vocal folds by the vocal tract. The vocal tract is everything from the nasal tract, tongue, teeth, lips, palate, etc. The particular configuration of the above organs (articulators) for every phoneme creates resonances at specific frequencies called formants. So, formants exist for both voiced and unvoiced sounds.

Pitch can be estimated by quantifying the period (using autocorrelation, say) or measuring the harmonics. Formant frequencies can be found by linear prediction analysis from the poles." (A.G. Ramakrishnan, 2013)

In 2019, A. S. Imran et al. [44] proposed text-independent speaker identification by using MFCC features and a deep neural network classifier. In their system, MFCC features of 3-second audio signals for each speaker were extracted and mapped into 299 x 13 2-D vector, whereas the value of '13' represents MFCC coefficient and '299' represents the dimension of the 2-D vector. Then each of these 2-D features was applied to a CNN algorithm. The CNN algorithm consisted of the following order: Two 2-D convolutional layers which each has a dimension of 3 x 3 in succession, a max-pooling which has a width of '2' and stride of '1', dropout layer that 25% of nodes used as dropout, 2-D convolutional layer, max-pooling layer, dropout layer, flatten layer, dropout layer, fully connected layer, fully connected layer, and classification layer. The system tested on 119 speakers in the MOOC database, and it achieved a 93.37% recognition rate for 3 seconds speech signal. It achieved a 94.44% recognition rate for 5 seconds speech signal, whereas 94.64% were achieved when 7 seconds speech was used.

The various research about ECG and Speech identification systems has been reviewed so far, and QRS complexes of ECG signal and MFCC features for speech signal have been decided as the most promising feature set. In the ECG recognition system, muscular noise, movement noise was stated as the cause of performance decreasing, whereas background and silence, in speech signals. For these reasons, various filters such as bandpass, smoothing, wavelet de-noising were applied in the pre-processing stage of their system. In speech processing, silence removing operation was applied on speech signals to increase the performance. Few researchers stated that vector quantization was applied to prevent the over-fitting problem. Based on this information, the proposed system took shape, and it will be presented in the following chapters.

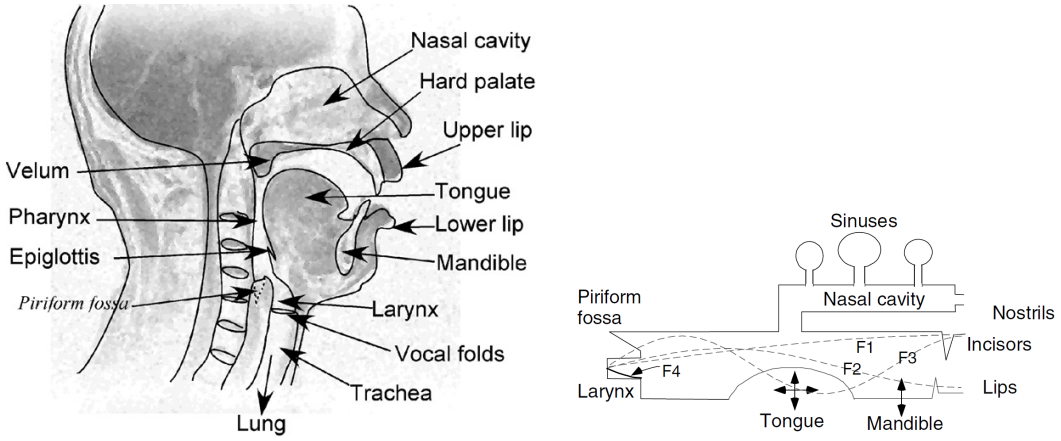
Chapter 2

Speaker Recognition System

2.1 Investigation of Speech Production

There are numbers of side branches in the vocal tract, which are the nose, piriform fossa, etc., and during speech production, these branches have less variation in some specific frequency regions. These frequency regions concerned with the side branches may be related to paralinguistic, which involves words in spoken language or non-paralinguistic information. The important point is the side branches have large variations across each speaker, whereas having small changes during speech production for the same person. They also do not change easily when the person has reached adulthood. However, the frequency regions of these side branches are not easily distinguishable. For these reasons, specific transformation is applied to extract acoustic features around those frequency regions which describe the individual information.

In order to understand intrinsic speaker features, we must know how and where the speaker features are encoded during speech production. Speech sound is formed when the source sound is passing through vocal tract filters. Although the vocal tract is treated as a single tube, it possesses a complicated shape that consists of the main tract and multiple side branches. In the figure 2.1(a), the human production system has been shown, and it can be seen that the vocal tract consists of complex side branches and cavities [45]. In the figure 2.1(b),



(a) Vocal tract of human production system

(b) Model of vocal tract with side branch

Figure 2.1: Human Speech Production [45]

the vocal tract is modeled, and the nasal passage is indicated as the biggest side branch. "In producing nasal and nasalized sounds, the nasal cavity is coupled with the oral cavity by lowering the velum. In some non-nasalized vowels such as /i/ and the voice bar of voiced stop consonants, a quite strong coupling takes place between the nasal and oral cavities via a transvelar coupling caused by the velum vibration" (Dang and Honda, 1994, 1997). Several paranasal cavities contribute anti-resonances to the transfer function of the vocal tract [45]. Because of the nasal cavity's complicated structure, nasal sounds do offer not only several distinguishing phonetic units but also contain individual information. In the figure 2.1(b), it can be seen that the piriform fossa is the entrance of the esophagus, and it is twin cone-like shaped. These side branches have anti-resonances from 4 to 5 kHz. The piriform fossa cavities are speaker-dependent and changeless during speech production. Much research has been established to clarify the correspondence of acoustical features to specific parts of the vocal tract. In these researches, it found that when producing vowels, the first three formants vary with the vocal tract while the fourth is almost constant. This phenomenon can be described by looking in the Figure 2.1(a) where the throat part of the vocal tract consists of the larynx and pharynx, and the larynx tube connects the pharynx via the outlet of the larynx. The larynx length is different for each speaker and contains high-frequency information. The vocal folds are also shown in Figure 2.1(a), locating

between the larynx and trachea, serving air input to lungs. It was estimated in the past research and, the fundamental frequency of the vocal fold's vibration is between 100 Hz and 400 Hz, depends on the length and stiffness of the vocal folds of each speaker. In summary, the speaker-specific features caused by different articulatory speech organs are distributed non-uniformly in low and high-frequency bands [45]. In the next sections, the MFCC feature extraction method, which focuses on both the low-frequency band and the high-frequency band of the vocal tract, will be reviewed.

2.2 Introduction to Speaker Identification

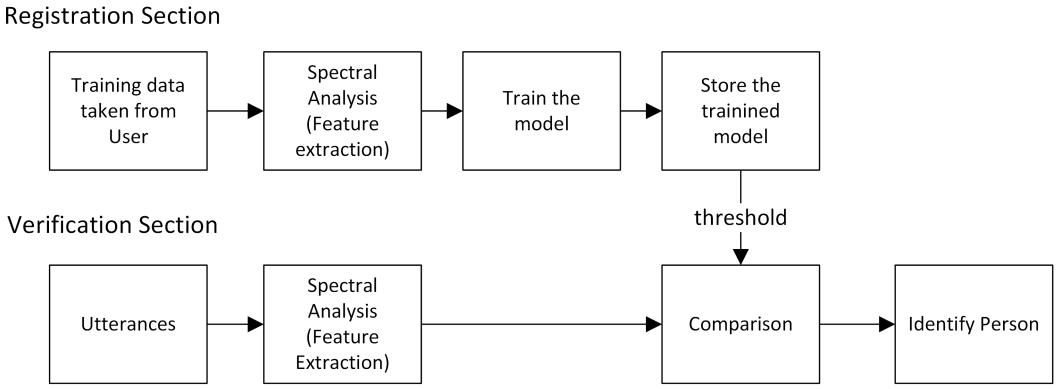


Figure 2.2: Basic Speaker Identification Block Diagram

Speaker Identification is a widely used biometric system, and it is the process of verifying the claimed identity of the registered speaker by using his/her voice characteristics. This type of system divides into two groups which are text-dependent and text-independent speaker identification. The difference is, the first one studies the keywords of the users while the last one studies the construction of the vocal tract of individuals. In Figure 2.2, the basic process of these two different speaker identification methods has been shown. In the registration phase, the speech signals of each speaker are given to the system. Then, significant features are extracted for each user by using spectral analysis and store as a training model. After constructing the training model, in the verification phase, spectral analysis is applied for the given unknown utterances, and feature of unknown

utterance is extracted. Then the decision is made by comparing the feature of unknown utterance with the features in the training model. The most similar feature in the training model to the unknown utterance is selected as the id of the person.

Chapter 3

Introduction to Biosignals

Human cells work like batteries, and these human cells produce small electricity from inside to outside of the ion concentrations of their membranes in the range of microvolts to millivolts. If the membrane got interrupted in any way, this disturbance would cause the rising of an action potential of that cell which eventually gets depolarized and repolarized. In Figure 3.1, the repolarization and depolarization of human cell are shown.

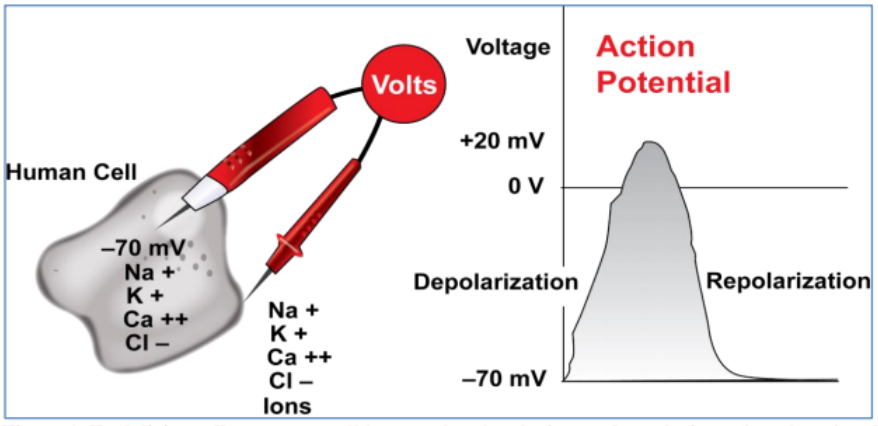


Figure 3.1: Polarization and depolarization in human cells [46]

3.1 Electrocardiogram

The depolarization and repolarization cycle of the heart can be seen in the figure 3.2. Electrocardiogram consists of 6 processes, and it can describe as follows.

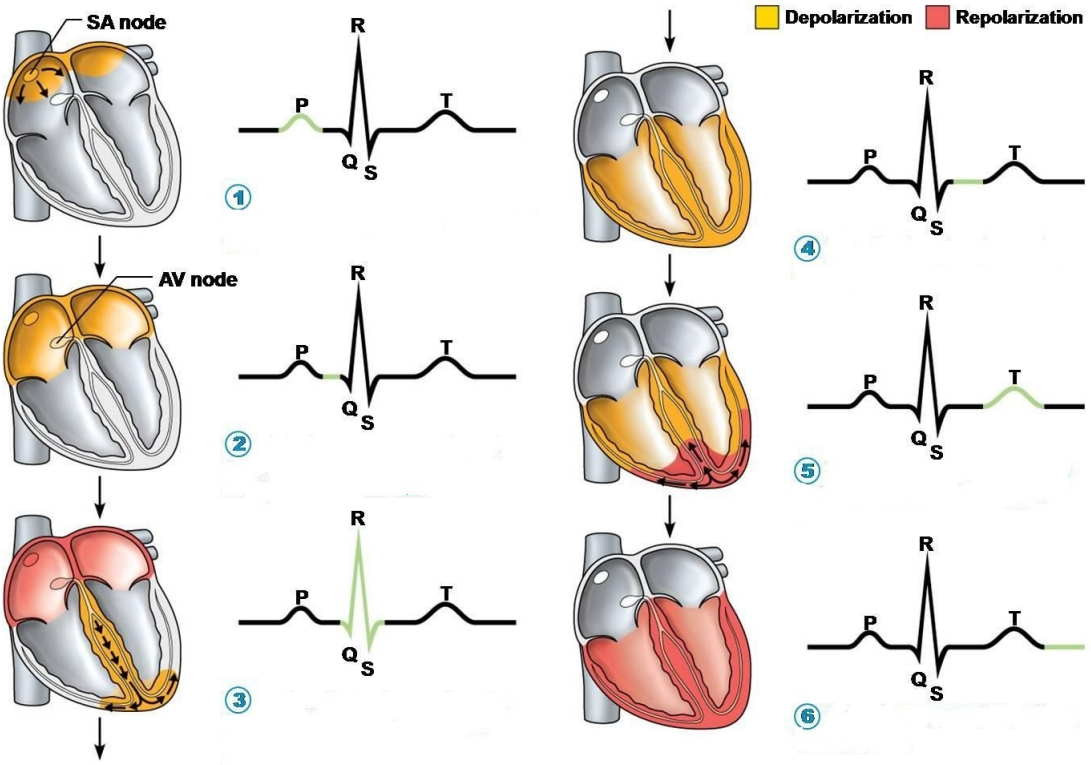


Figure 3.2: Depolarization and Repolarization in hearth cycle [46]

1. Atrial depolarization, which is initiated by the SA node, causes the P wave
2. When atrial depolarization is complete, the impulse is delayed at the AV node
3. Ventricular depolarization begins at the apex and causes the QRS complex. Then Atrial repolarization occurs
4. Ventricular depolarization is complete
5. Ventricular repolarization begins at the apex and causes the T wave
6. Ventricular repolarization is complete

3.2 Electrocardiogram Measurement

The electrical activity of the heart produces currents that spread from surrounding tissue to the skin. By attaching electrodes to the skin, electric current activity

can be easily detected. Then these electrical activities, which represent the depolarization and repolarization cycles of heath, turn into graphical waveform by a differential amplifier. The issue is, electric currents are in the range of millivolts to microvolts. In order to increase the accuracy of measurement, these current activities are applied to the instrumentation amplifier so that they can bring to the logical voltage level. The other issue is, these electrical activities radiate from the heart to the skin in many directions. For this reason, multiple electrodes are placed onto the human body so that the total picture of the human heart can be represented. The ECG records which represent the different perspectives are called **leads**, whereas recording of the electrical activity of the heart called **electrocardiogram** [47]. Each lead provides the electrical activity of the heart, and they can be measured between two points that are attached in opposite directions of the heart. The magnitude of the waveform, which is recorded from these two points, will change if the current of the heart and its direction to the skin change.

3.2.1 12-Lead ECG Measurement

12-Lead measurement setup provides 12 different views of electric activity of the heart. By placing 9 electrodes on the patient's limbs and chest, 12 waveforms of the heart activity are obtained at different angles and planes. The measurement points of each lead can be described as

Lead-I: Right Arm $(-)$ and Left Arm $(+)$, whereas Right Leg is reference point. Lead-I observes the heart "from the left" because of placing a positive probe to the left arm, and it is described as observation of 0^0 angle of heart activity in the vertical plane

Lead-II: Right Arm $(-)$ and Left Leg $(+)$, whereas Right Leg is reference point. Because of exploring electrode places onto Left Leg, it is described as observation of 60^0 angle of heart activity in the vertical plane

Led-III: Left Arm (−) and Left Leg (+), whereas Right Leg is reference point. It observes the heart from an angle of 120^0 in the vertical plane

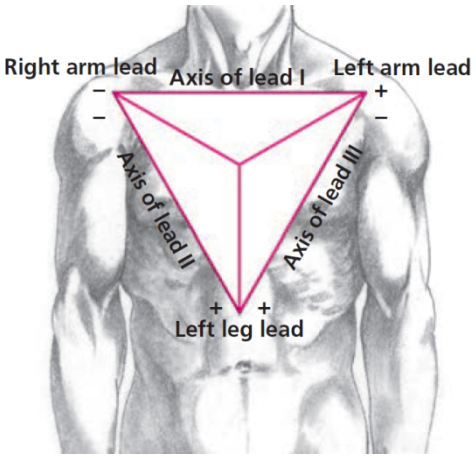


Figure 3.3: Einthoven's Triangle [47]

The measurement point provides information about the heart's frontal plane forms like a triangle. Axes of the three limbs, which are lead I, II, and III, are known as Einthoven's triangle lead.

Lead aVL: It is the combination of Lead-I and Lead-III and formulated as

$aVL = \frac{\text{Lead-I} - \text{Lead-III}}{2}$ (Goldberger's lead system aVL) and it observes the heart from an angle of -30^0 in the vertical plane.

Lead aVR: It is the combination of Lead-I and Lead-II and formulated as

$-aVR = \frac{\text{Lead-I} - \text{Lead-II}}{2}$ (Goldberger's lead system aVR) and it observes the heart from an angle of 30^0 in the vertical plane.

Lead aVF: It is the combination of Lead-II and Lead-III and formulated as

$aVF = \frac{\text{Lead-II} - \text{Lead-III}}{2}$ (Goldberger's lead system aVF) and it observes the heart from an angle of 90^0 in the vertical plane.

Lead V1: It is the combination of $V_w(-)$ and $V_1(+)$ point. V_1 electrode is in the position of the fourth intercostal space to the right sternum. It observes the heart from an angle of 0^0 in the horizontal plane.

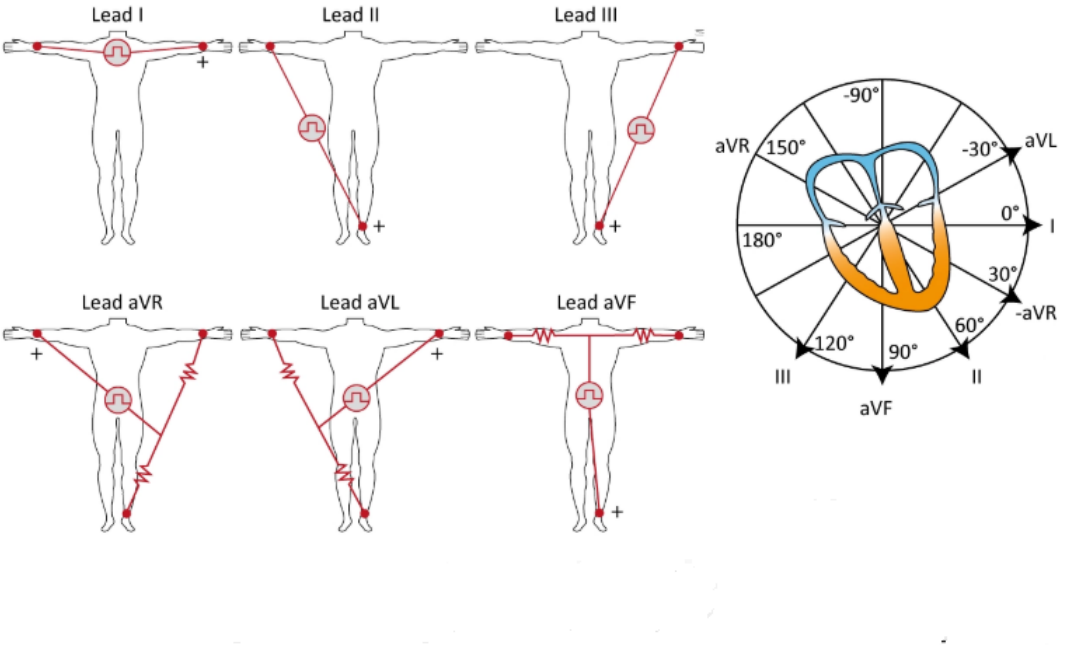


Figure 3.4: Replacements of Einthoven's Lead, and Goldberger's Lead and their review of the heart's electric activity in the vertical plane [48]

Lead V2: It is the combination of $V_w(-)$ and $V_2(+)$ point. V_2 electrode is in the position of the fourth intercostal space to the left sternum. Lead V1 and V2 are known as "septal leads", and they observe the ventricular septum but may occasionally show the changing originated from the right ventricle. It observes the heart from an angle of 30^0 in the horizontal plane.

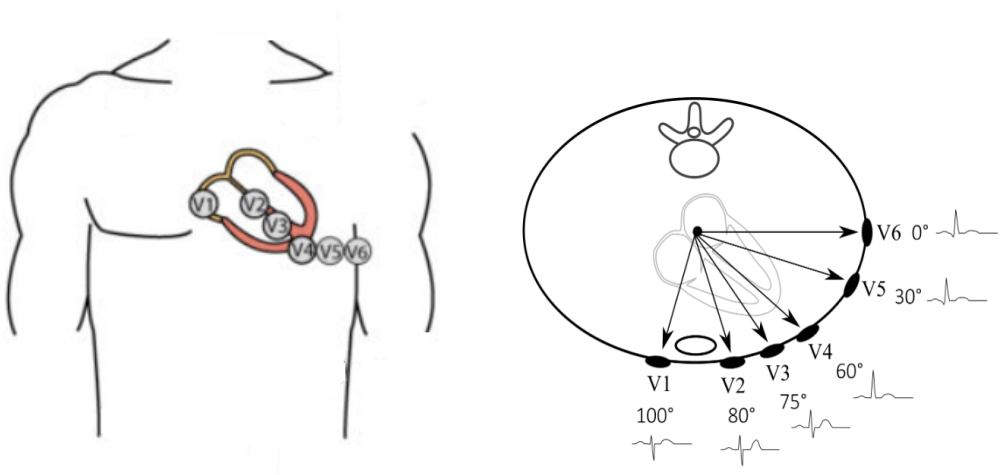
Lead V3: It is the combination of $V_w(-)$ and $V_3(+)$ point. V_3 electrode is in the position under V2 and V4 points and diagonally placed between V2 and V4. It observes the heart from an angle of 60^0 in the horizontal plane.

Lead V4: It is the combination of $V_w(-)$ and $V_4(+)$ point. V_4 electrode is in the position between rib 5 and 6 in the midclavicular line. It observes the heart from an angle of 75^0 in the horizontal plane.

Lead V5: It is the combination of $V_w(-)$ and $V_5(+)$ point. V_5 electrode is in the position same as V4 but in the anterior axillary line. It observes the heart from an angle of 80^0 in the horizontal plane.

Lead V6: It is the combination of $V_w(-)$ and $V_6(+)$ point. V_6 electrode is in the position same as V4 and V5 but in the midaxillary line. It observes the heart from an angle of 100^0 in the horizontal plane.

,whereas $V_w = \frac{1}{3}(RA + LA + LL)$ and Right Leg is reference point.



(a) Replacement of Wilson's Lead

(b) Review of the heart's electric activity in the horizontal plane

Figure 3.5: Replacement of Wilson's Lead and their review of the heart's electric activity in the horizontal plane [48]

Lead-I, II, and III are known as Einthoven's Leads, aVL, aVR, and aVF are known as Golberger's Leads, and provide information about the frontal plane of the heart, whereas V_1, V_2, V_3, V_4, V_5 and V_6 Leads are known as Wilson's Leads and provide the information of the horizontal plane of the heart [47, 48].

3.2.2 Common Monitoring Problems

Waveforms shown in the figures below illustrate the most common problem when recording and monitoring the ECG signals.

3.2.2.1 Baseline Wandering

Movements of patients' restlessness or poor quality of electrodes placed on the skin of patients cause such problems. It causes the baseline of ECG signals to change over time.

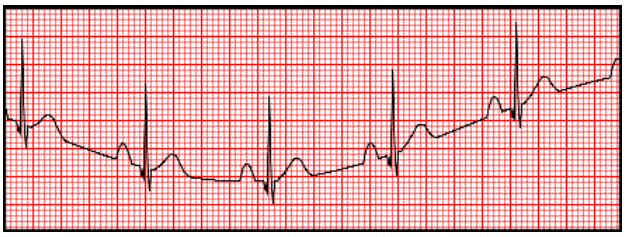


Figure 3.6: Baseline wanders in ECG signal [47]

To solve the problem, 0.5 to 1 Hz, high pass filter should be applied onto the ECG signals, which have baseline wanders. It can also be reduced by encouraging the patient to relax and by replacing or repositioning the electrodes properly.

3.2.2.2 Power line Interference

The United States provides their citizens alternative current, which has a frequency of 60 Hz, whereas Europe provides 50 Hz. Power line interference is seen when a power supply of the ECG circuit is poorly grounded or not having filters to filter out specific power noise. The corrupted ECG signal is seen in figure 3.7 where has additional frequency on the original signal causes ripples.

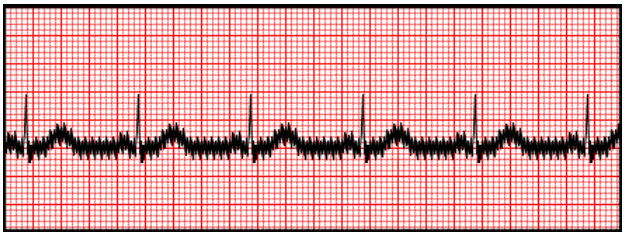


Figure 3.7: ECG signal which has AC interference [47]

Power line interference can be eliminated by applying band stop or notch filters. Filter specifications must be calculated by considering the frequency of the power line the countries provided.

3.2.2.3 Muscle Tremor

During measurement of the ECG signals, the patient's muscular signals will also be recorded if the patients prop themselves up by their arms or patients are cold and shivering. This type of noise is called muscle tremor, and it can be seen in the figure 3.8. This type of noise is not easy to filter out because muscle noise does not have a specific frequency range and can be occurred in the range ECG signal's frequency. For this reason, de-noising filters (transforms) such as wavelet, EMD, or smoothing filter are applied to reduce the muscle noise in exchange for losing some of the ECG information.

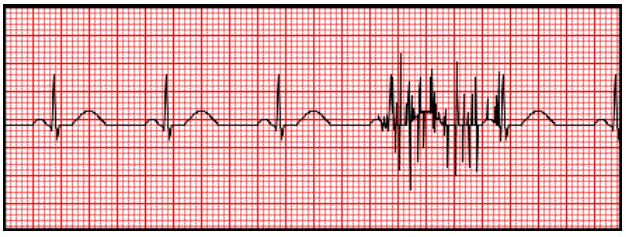


Figure 3.8: Muscle interference in ECG signal [47]

3.2.2.4 Misleaded Electrodes

Attention should be paid while placing electrodes because the direction of the current flow changes whenever the polarization of the ECG signal changes.

Figure 3.9: Reversed lead ECG record [47]

3.3 Components of an ECG Waveform

An ECG complex represents the electrical activity of the heart in one cardiac cycle. ECG signals consist of three waveforms in one cardiac cycle; QRS complex, P wave, and T wave. Additional to these waves, u wave can sometimes be seen.

The electrical activity can also be separated into segments and intervals, which are the ST segment, the PR interval, and the QT interval. In the figure 3.10, J point marks the end of the QRS complex and the beginning of the ST segment [47].

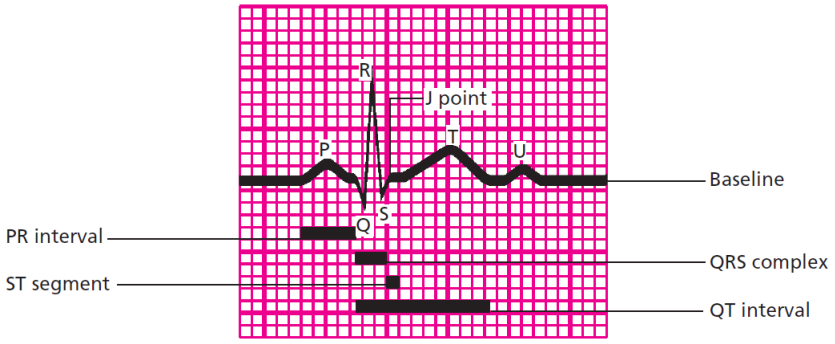


Figure 3.10: Components of ECG waveform [47]

In the figure 3.11, the page grid illustrates the vertical, horizontal axis and their measurement values.

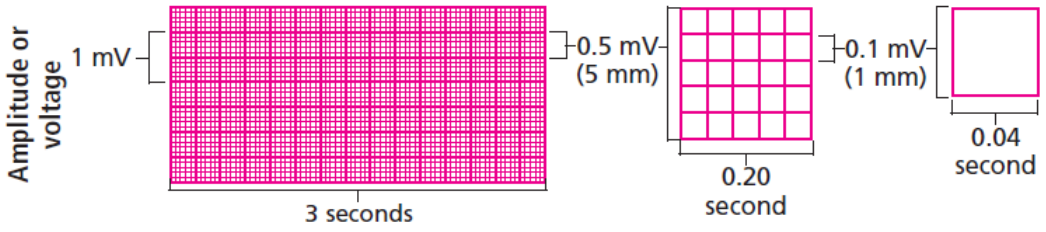


Figure 3.11: The grid for ECG signal [47]

3.3.1 P wave

The P wave is the first component of a normal ECG signal. It represents atrial depolarization or transmission of an electrical impulse through the atria [47]. The P wave characteristics of normal ECG are given as

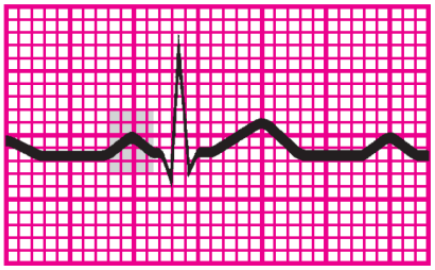


Figure 3.12: P wave [47]

- Amplitude: between 0.2 to 0.3 mV
- Duration: between 0.06 to 0.12 sec.
- Location: before QRS complex
- Deviation: positive in Lead I, II, aVF and V_2 to V_6, negative in other leads.

3.3.2 PR Interval

PR interval follows the atrial impulse from the atria to the AV node, the bundle of this, right and left branches [47]. The PR interval characteristics of a normal ECG are given as

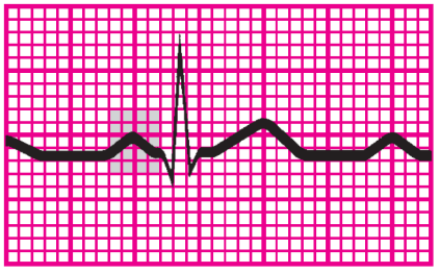


Figure 3.13: PR Interval [47]

- Duration: between 0.12 to 0.20 sec.
- Location: It is between the beginning of the P wave to the beginning of the QRS complex

3.3.3 QRS Complex

QRS complex tracks the P wave, and it represents the depolarization of ventricles. After the ventricles depolarize, the blood is ejected from the ventricles and is pumped through the arteries [47]. QRS complex characteristics of a normal ECG are given as

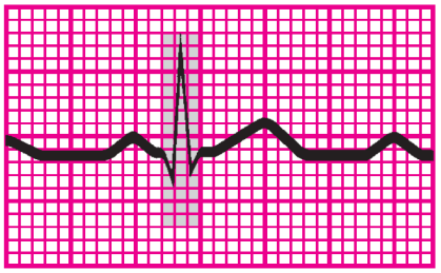


Figure 3.14: QRS Complex [47]

- Amplitude: between 0.5 mV to 3 mV, however it differs for each leads
- Duration: between 0.06 to 0.10 sec.
- Location: It tracks the PR interval (between the beginning of the Q wave to end of S wave)
- Deviation: positive in Lead I, II, III, aVL, aVF, and V_4 to V_6, biphasic in V_3, negative in other leads

3.3.4 ST Segment

ST-segment represents the ventricular depolarization and the beginning of ventricular repolarization [47]. ST-segment characteristics of a normal ECG are given as

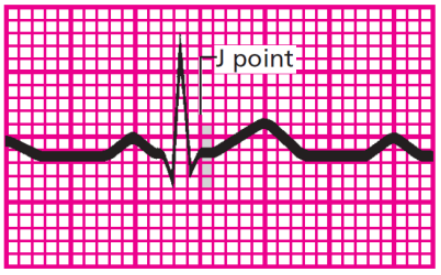


Figure 3.15: ST Segment [47]

- Location: It spreads from S wave to the beginning of T wave
- Deviation: usually on baseline, may change between -0.05 mV to 0.1 mV in some leads

3.3.5 T Wave

T wave represents the ventricular recovery, or period of repolarization [47]. T wave characteristics of a normal ECG are given as

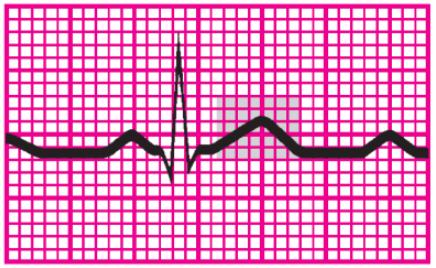


Figure 3.16: T Wave [47]

- Amplitude: vary between 0.05 mV to 0.5 mV
- Location: It tracks the ST segment
- Deviation: positive in Lead I, II, and V_2 to V_6 , negative in aVR ,vary in leads III, and V_1

3.3.6 QT interval

QT interval is the time needed for ventricular depolarization and repolarization. QT interval changes according to heart rate. QT interval is shorter when heart rate is increased [47]. QT interval characteristics of a normal ECG are given as

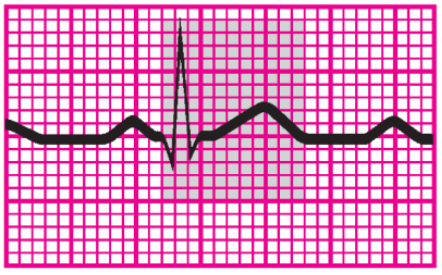


Figure 3.17: QT Interval [47]

- Location: It spreads from the beginning of QRS complex to the end of T wave
- Duration: It varies according to gender, age, and heart rate. It is between 0.36 to 0.44 sec.

3.4 Introduction to ECG based Person Identification System

A typical form of ECG-based Identification system is illustrated in figure 3.18, and the pre-processing block and multiple filters are applied to the signals to increase the performance of the system. The other processes are nearly the same as speech-based identification systems, as previously mentioned in Chapter 2.

In an ECG-based system, feature extraction methods are separated into two groups which are fiducial based and non-fiducial based points, and shown in the figure 3.19. In the fiducial-based feature extraction, wave peaks, boundaries, slopes of ECG signals are being used as features and categorized into three groups which are temporal, amplitude, and morphological. Temporal relation between

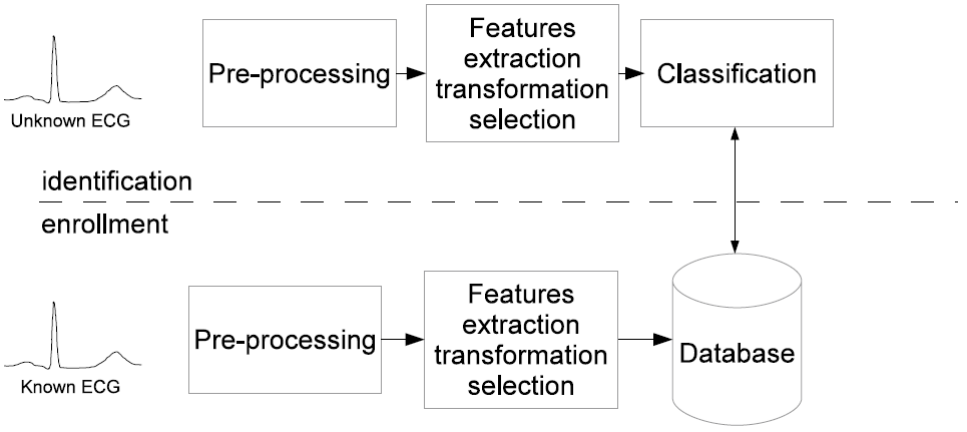


Figure 3.18: Typical ECG-based Person Identification System Block diagram [49]

various ECG waves such as P wave, QRS complex, and T wave reflects the stage of heart's stimulation along its electrical paths and can be used as a feature for identification. P wave duration, QRS complex duration, T wave duration, PR segment, ST segment, PR interval, QT interval can be given as an example to the temporal features. The amplitude of the ECG waves can be easily recognized and used as features which are P wave amplitude, Q, R, S point amplitude, T wave amplitude. Morphological features are a bit different from temporal and amplitude features, and it carries the information of ECG shapes, either as whole or partial waves. In most morphological features extraction, slope among waves such as ST, RS segment slopes, and angles described by Q, R, S waves have primarily been used. Polynomial expansions which synthesize the heart morphology or Hermite interpolation coefficients derived by fitting QRS complex are mostly among the morphological-based features. These types of features require accurate detection of fiducial, and results achieved by this kind of feature are dependent on how clearly these points are being found. For this reason, a new approach that does not require fiducial point has been reported to overcome this problem and is called non-fiducial-based features. All the techniques described on non-fiducial-based features are based on the assumption of ECG signal's highly repetitiveness and categorized into three groups. In autocorrelation-based features, normalized autocorrelation coefficients are calculated by using randomly selected ECG signals that have a specific length in time, such as 5 seconds or 10 seconds, and used

as a feature. In the phase space-based approach, ECG signals are characterized into two dimensional or even three-dimensional space by using the time delay technique so that they can highlight unexplored peculiarities of cardiac activity. In the frequency-based approach, the frequency content of ECG signals is examined by using spectrum transformation methods such as LPC, MFCC, EMD, or Hilbert-Huang transforms and are used as feature set [49].

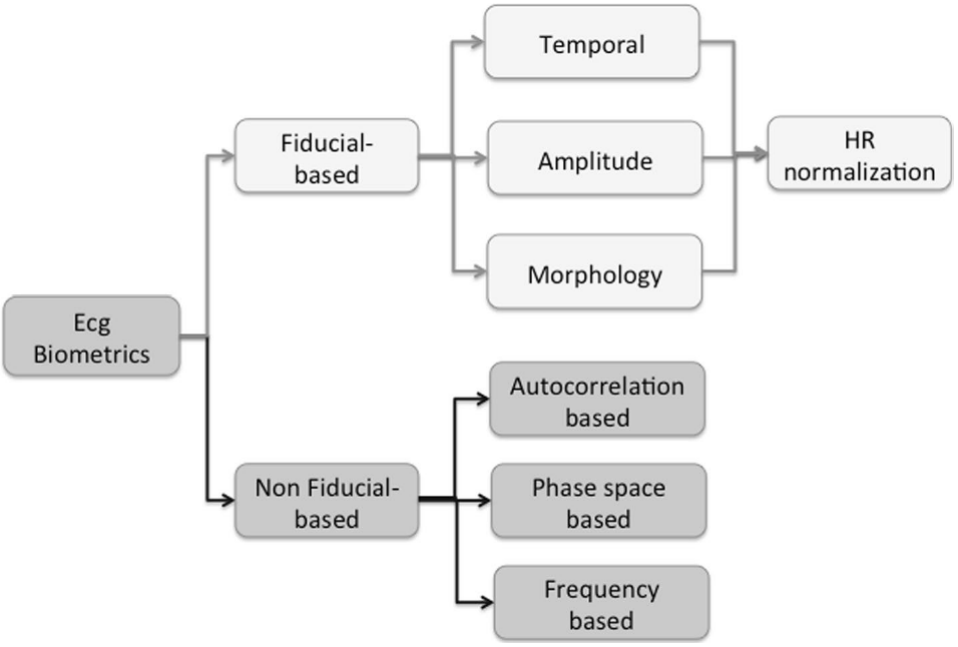


Figure 3.19: Feature Extraction Methods for ECG-based Person Identification System [49]

Chapter 4

Design of Speech and Fingertip ECG Measurement System

4.1 Block diagram of Speech and ECG Measurement System

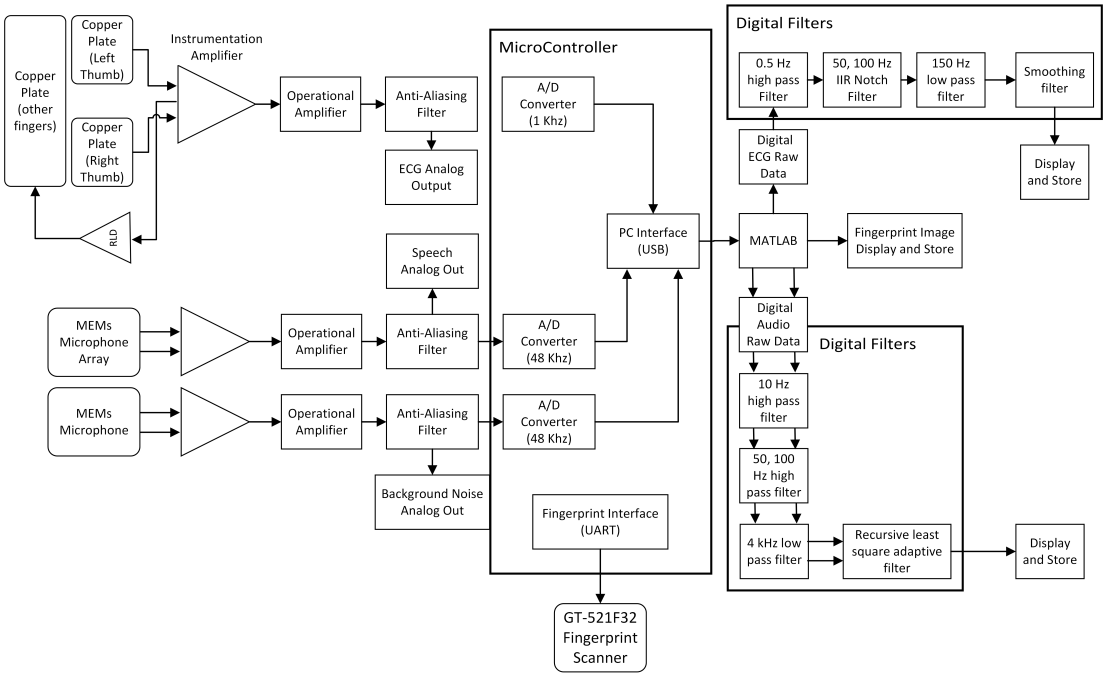


Figure 4.1: Block Diagram of Speech and ECG Measurement System

Traditional ECG measurement system works with placing multiple Ag/AgCl electrodes on the human body by considering the current direction of the heart. Although this type of system measures the heart activity accurately with low muscle

interference, it takes time to place the electrodes to the patient's body which indicates that it is not easy to use. For this reason, a new type of ECG measurement system is needed, which both will stop consuming the user's effort and time and will only extract the necessary and significant current activity of the heart for the identification system. Block diagram of the proposed system is given in figure 4.1. It consists of two independent analog structures in the same case that do not affect each other's working but are using the same Micro Controller.

The first analog structure is used to measure ECG signal from Lead-I connection points, whereas the second structure is used to record speech signals. In the ECG measurement part, the first two copper plates are used for patients to hold it with both left and right thumbs while pressing the third copper plate with their remaining fingers. Action potentials in the range of millivolts are captured from these two copper plates by instrumentation amplifier, and ECG signal is formed by differentiating these two action potentials over time while the third copper plate is used as negative feedback to increase the Common Mode Rejection Ratio (CMRR) because high CMRR must be so to amplifies the small signals, accurately. For the safety of patients and proposed circuits both, electrostatic discharge (ESD) protection and high input resistors have been used to prevent current leakage. After the instrumentation amplifier, the operational amplifier is used to increase the overall amplitude of ECG signals in the range of millivolts to the logic voltage value. Therefore, instrumentation and operational amplifier's gain are adjusted until the total gain approaches 1000. Then anti-aliasing filter (low-pass filter) is used for eliminating the high-frequency component of ECG before the analog-digital converter so that high-frequency components can not disrupt the desired ECG signal. The ideal low pass filter cut-off frequency is determined by looking at the sampling rate of the analog-digital converter. Because of having a 1000 Hz sampling rate, the high-frequency component on the ECG signal must be lower than 500 Hz to obtain the digital signal perfectly. For this reason, the cut-off frequency of the low pass filter is determined as 159 Hz. In addition to digital output, it offers analog ECG output after an anti-aliasing filter. PC Interface

protocol is established and developed in C language by using Keil Embedded Development Tool for ARM microcontroller, which is STMF407VG. The protocol is separated into two parts and can be changed by pressing the button in 3 seconds. The first interaction provides both visualization and registration of ECG or Speech data in the Oscilloscope Open Source Application. Pressing the button shortly gives us access to which data (Speech or ECG) is wanted to record and visualize. The second interaction provides direct Speech and ECG data access on the computer simultaneously. Because the Matlab interface does not provide us real-time plotting, this type of solution was realized. In summary, the first interaction is for creating an ECG and speech database to training the system, whereas the second interaction is for testing the proposed system given ECG and speech data simultaneously within a certain period of time. For future works, fingerprint interface with PC by using Matlab is also established, and it is ready to capture and transform the fingerprint data to RGB or grayscale image. The system offers noise elimination on both recorded ECG and Speech data. 4^{th} order 0.5 Hz IIR Butterworth high pass filter, 150 Hz window-based FIR Low pass filter which the window function was hamming and window length was 48, 50 and 100 Hz 2^{th} order IIR notch filters are applied on ECG data whereas 10 Hz 4^{th} order IIR Butterworth high pass filter, 4 kHz window-based FIR low pass filter which the window function was hamming and window length was 48, 50 and 100 Hz IIR notch filters are applied on Speech data. At the end of the filtered ECG signal, a smoothing operation is applied to the signal to reduce the muscle noise by considering losing the high-frequency component of the ECG signal. The length of the smoothing filter is determined as 11.

The second analog structure is established to record two Speech signals where one will be used as the original signal (recorded by microphone array, for the purpose of a directional microphone), the other will be used as noisy speech which closes to the noise environment (recorded by omnidirectional microphone). The differential output port of these microphones is used to minimize external interference. 16 MEMs microphone is aligned into a circular shape with taking into account their

distances to each other to establish a directional voice record system. Other than instrumentation amplifier structure, anti-aliasing filter cut-off frequency, analog-digital converters sampling rate, and in addition of adaptive RLS algorithm in filters operations, it is the same as the first analog structure. However, in the second analog structure, two speech signals are extracted and used to decrease, if possible, eliminate the background and microphone internal noises. Because the characteristic of MEMs microphones and its circuits are the same in both omnidirectional and directional recordings, RLS algorithm does reduce not only the background noise but also eliminate the noise caused by circuit or microphone itself.

4.2 Schematic of Speech and ECG Measurement System

In figure 4.3, a schematic of ECG analog structure is illustrated. ECG circuit is designed as both single supply mode where the device fed with 3.7 V regulator which powered by USB and dual supply volt where the device fed with +/- symmetric battery supply rail. The device is also designed to work with only 2 fingers where only the left and right fingers touch the copper plate. It is convenient and easy to use with a single supply mode or 2 fingers measurement, but both cause the power line noise to increase, especially using 2 finger mode. An INA333 instrumentation amplifier is used to differentiate the signal from the left and right side of the fingers caused by the activity of the heart. “R5” and “R6” are the gain resistors for the instrumentation amplifier, and they are calculated to achieve the gain of 51. An integrator amplifier which is feed with the output of the instrumentation amplifier, is given to the reference point of the instrumentation amplifier to reduce the baseline wander as much as possible. Then raw ECG signal passes through the first order low pass filter, which has a 159 Hz cut-off frequency and gain of 20. This operational amplifier both increases the signal into the range of 1 V to 5 V and eliminates the frequency components higher than 159 Hz. ECG analog output is provided by a 3.5mm audio port (J1) and ECG

socket (P11). This also gives us the option of recording the ECG data through the audio port on the PC.

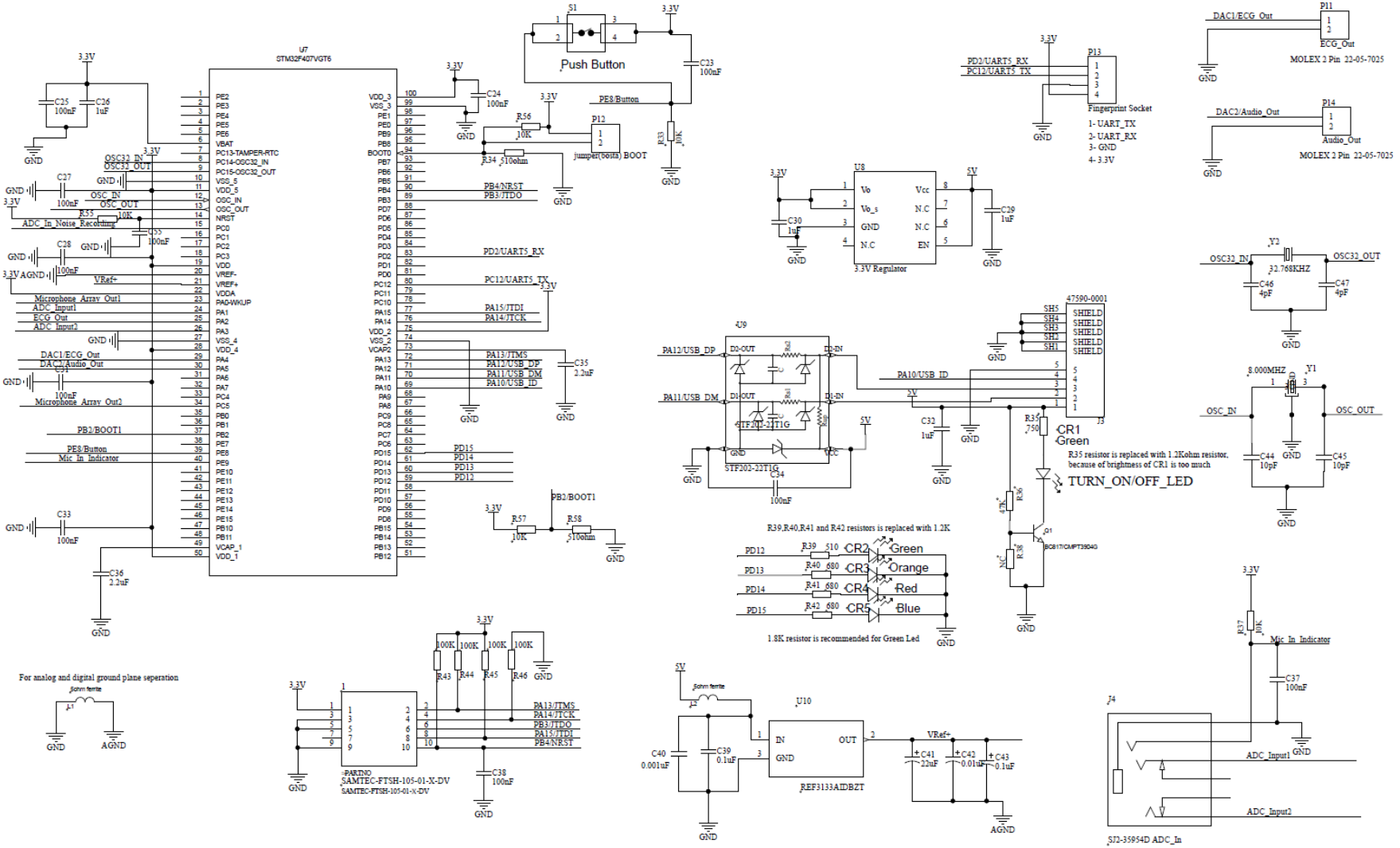


Figure 4.2: Circuit of Speech and ECG Measurement System (MCU Control)

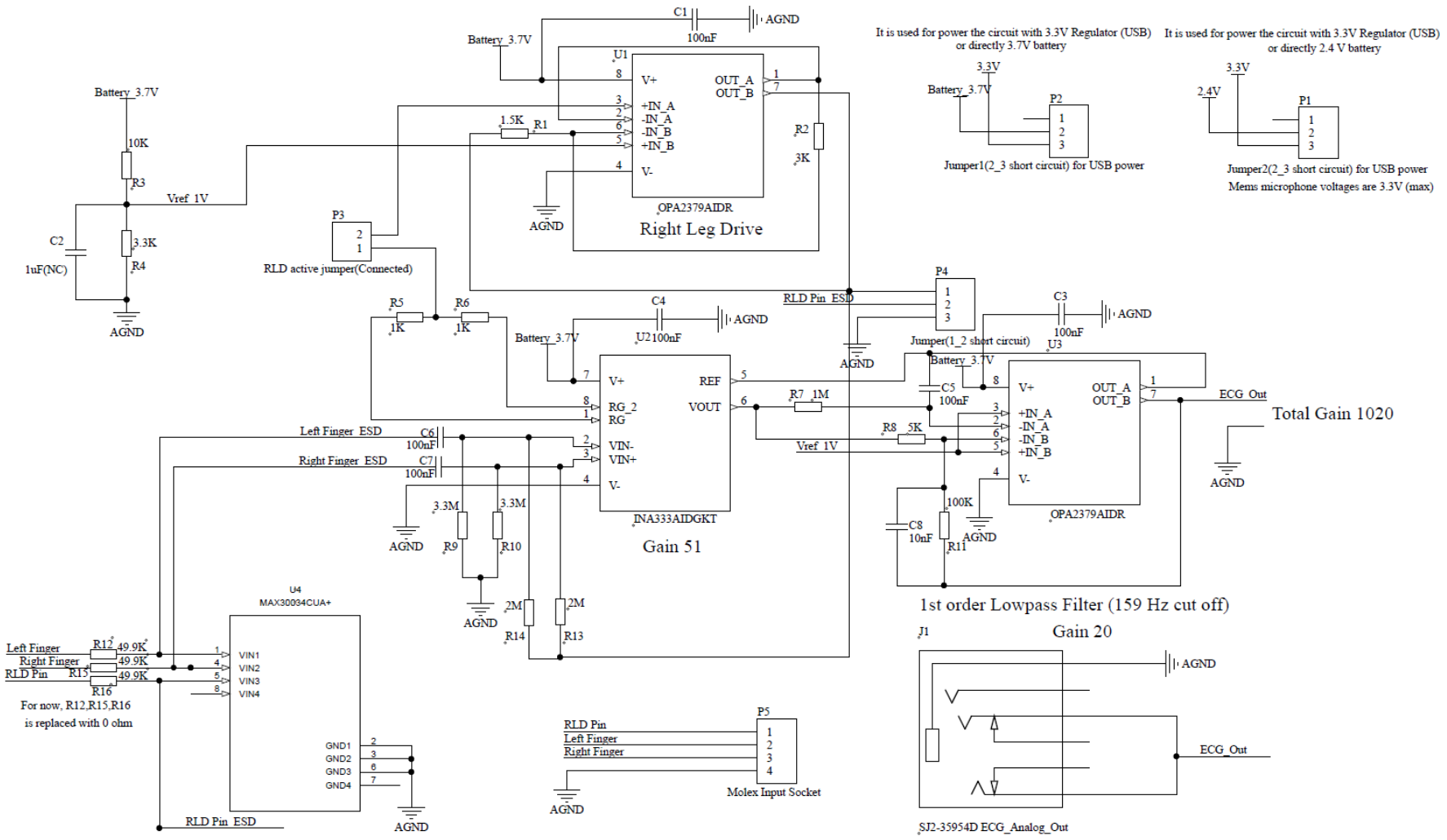


Figure 4.3: Circuit of Speech and ECG Measurement System (ECG Analog Structure)

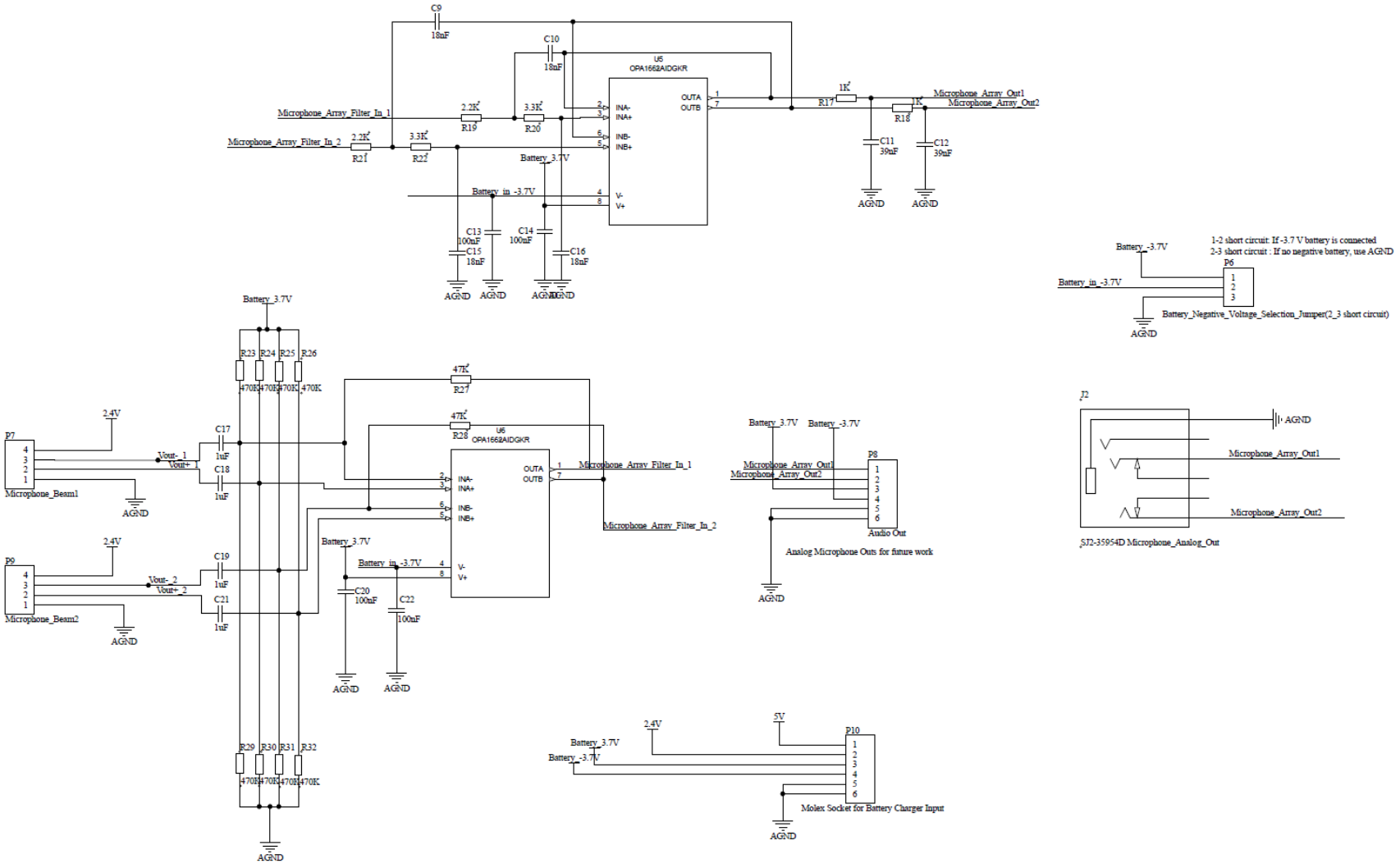


Figure 4.4: Circuit of Speech and ECG Measurement System (Speech Analog Structure)

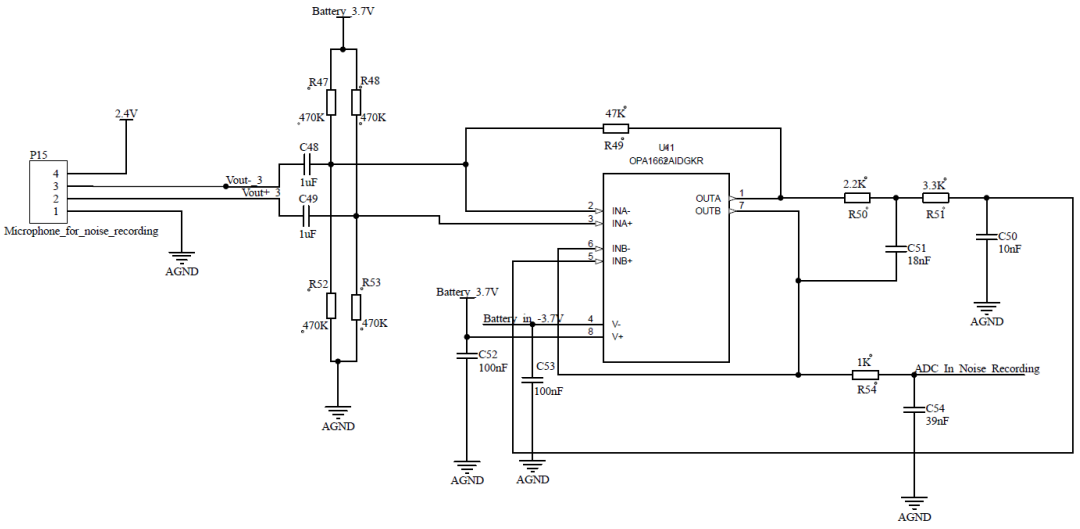


Figure 4.5: Circuit of Speech and ECG Measurement System (Noise Recording Structure)

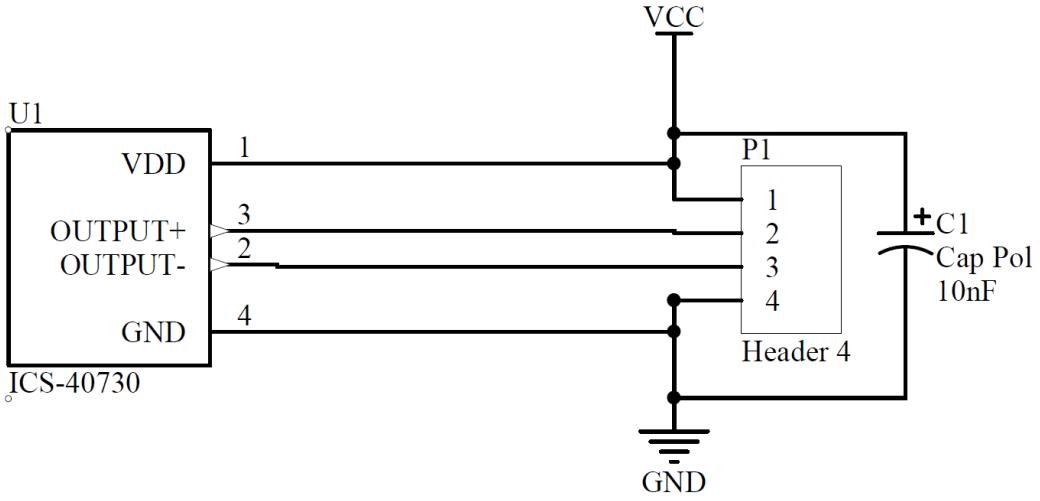


Figure 4.6: MEMs microphone module for noise and microphone array construction

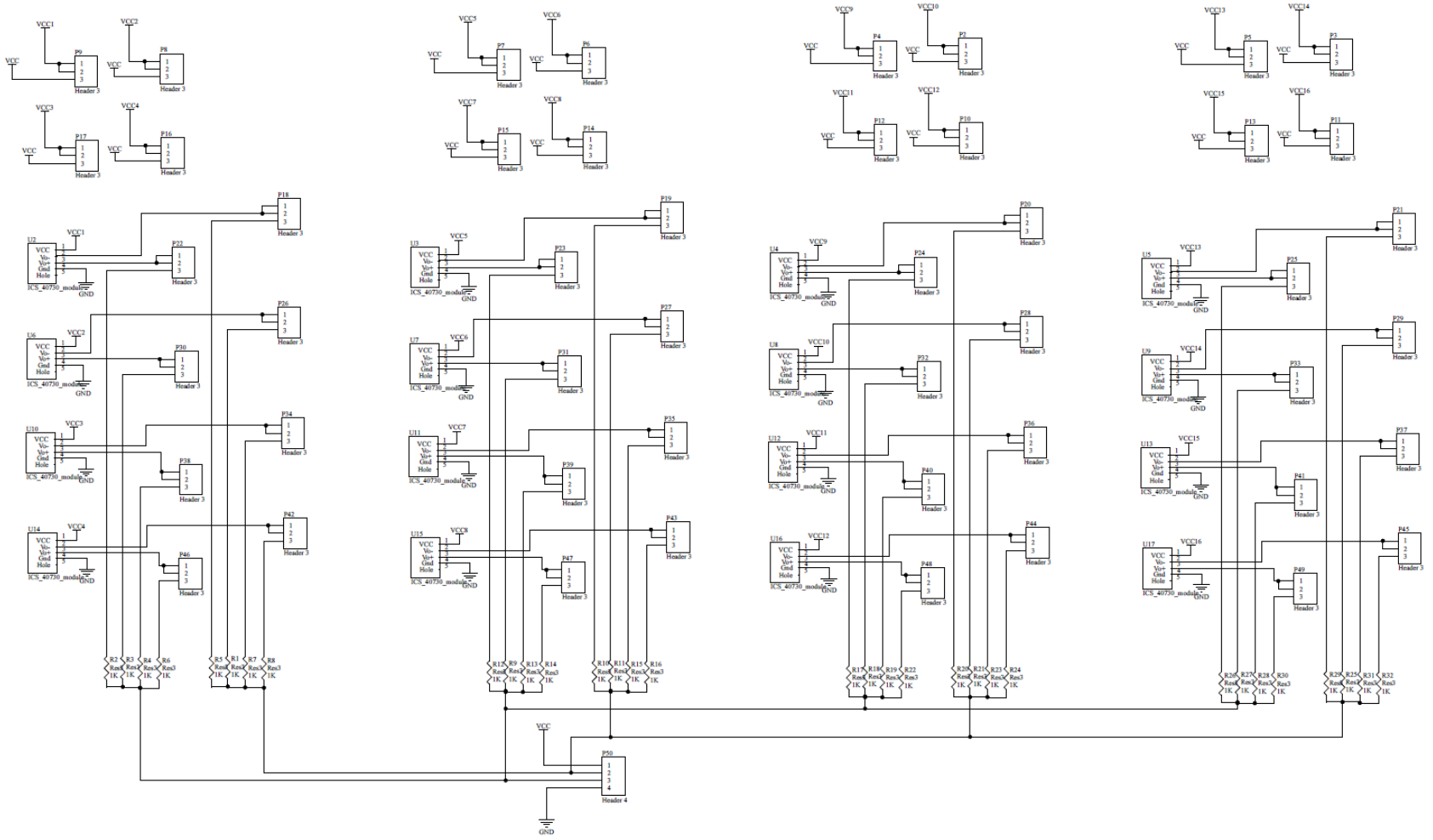


Figure 4.7: Microphone Array Beam Module (16 MEMs microphone) for directional speech recording

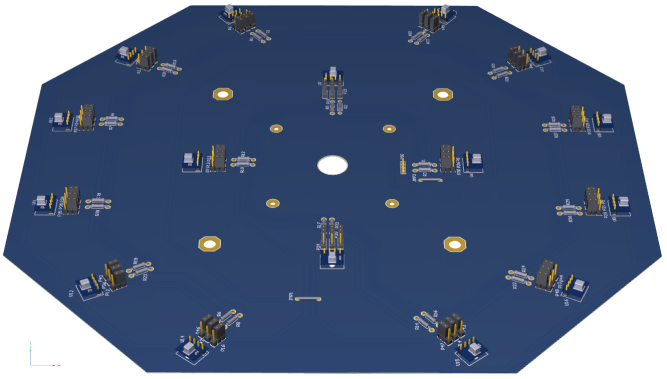


Figure 4.8: 3D model of Microphone Array Beam

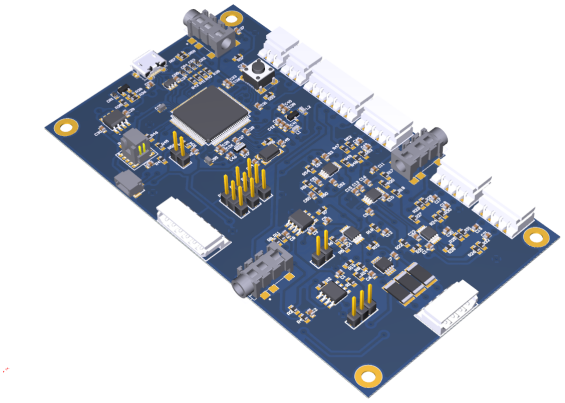
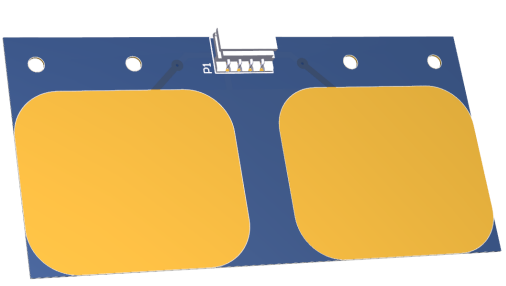
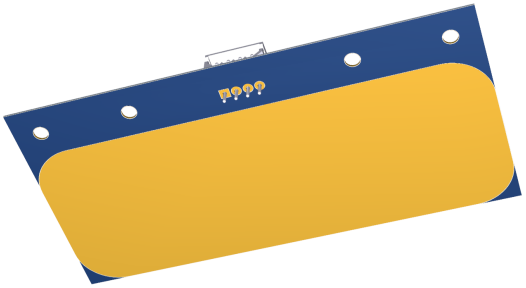


Figure 4.9: 3D model of Speech and ECG measurement system



(a) 3D model of Fingertip Connector Front plane



(b) 3D model of Fingertip Connector Back plane

Figure 4.10: Fingertip ECG Connector

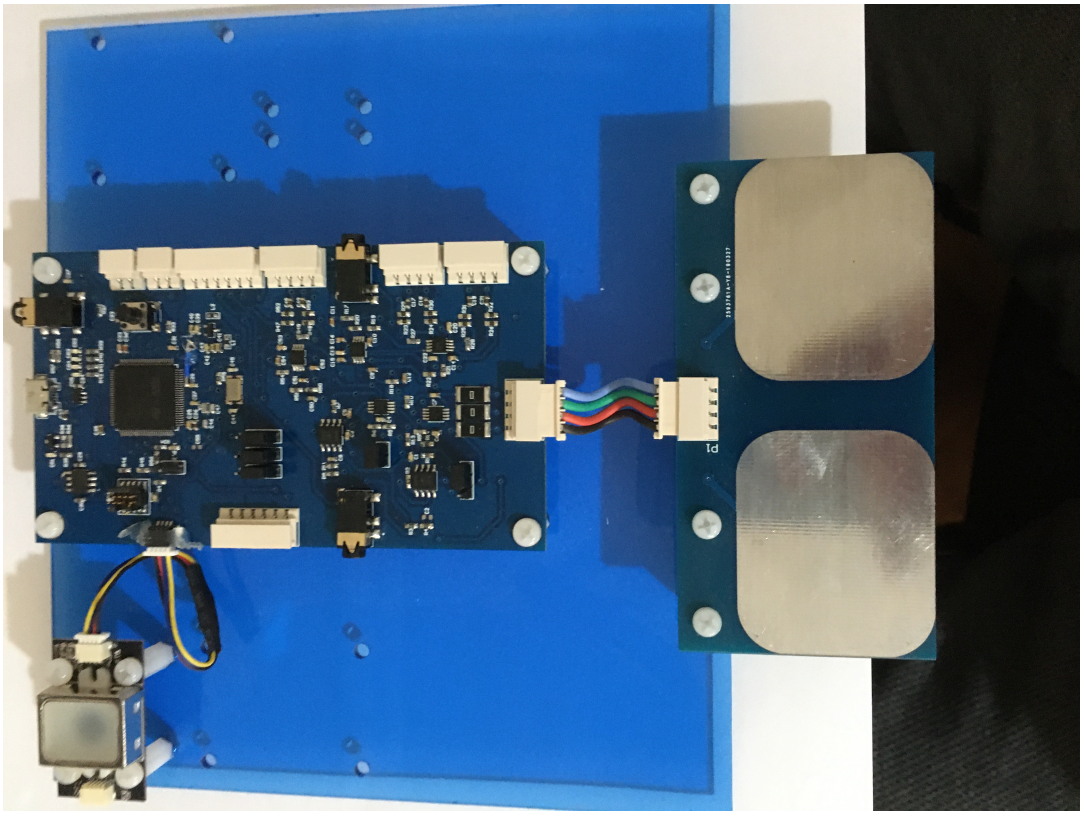


Figure 4.11: Speech and ECG circuit implementation

In figure 4.2, the schematic of the MicroController unit is illustrated. STM32F407 microcontroller is used to transfer the speech and ECG data signals to PC through a USB port. It is a 32-bit ARM-Cortex M4 core-based microcontroller that has a speed of 168 Mhz. STF202 (U9) EMI filter is used to suppress interference generated by the device, other equipment and to protect the device from electromagnetic interference signals present in the environment because electromagnetic interference (EMI), or radio frequency interference (RFI) are types of electronic emission which impair the circuit performance. Protecting against the effect of electrostatic discharge, MAX30034 (U4) ESD protection is an essential requirement, and it is used in the area where patients contact with the circuit. It protects both the device's reliability and patient safety.

In figure 4.4, the schematic of speech analog structure is illustrated. MEMs microphone arrays are connected to P7 and P9 sockets. The first intention is to use two microphone arrays putting back to back to create an end-fire array

topology and by delaying and summing the speech signals by considering the distance differences to increase the directionality. Later on, one of the ports is disregarded, and only one microphone array which has broad-side array topology is used. The first operational amplifier is used to differentiate the positive and negative audio signals for eliminating the noise on the cable to obtain a raw speech signal. Then second order low pass filter, which has a 4 kHz cut-off frequency, is applied onto the raw speech signal to eliminate high-frequency components. After that, the filtered speech signal is converted to a digital signal through an analog-digital converter which has a sampling rate of 48 kHz.

In figure 4.5, the noise recording circuit is illustrated. The architecture of the circuit and component's value is the same as figure 4.4, so that characteristic of directional and omnidirectional microphone become the same. This circuit is used to record the background noise and noise caused by the circuit and eliminating them by using RLS adaptive algorithm.

In figure 4.7, MEMs microphones are replaced with a distance to each other to achieve broad-side array topology, and this will be reviewed in the next chapters. In figure 25-a, the MEMs microphone schematic is illustrated. ICS-40730 MEMs microphone, which is an ultra-low noise, has a differential analog output, and has bottom ported architecture, is used to construct microphone array. Its 74 dB SNR +/- 2 dB sensitivity tolerance makes it an excellent choice for far-field voice application [50].

In summary, the properties of the speech and ECG measurement system can be explained in figure 4.12.

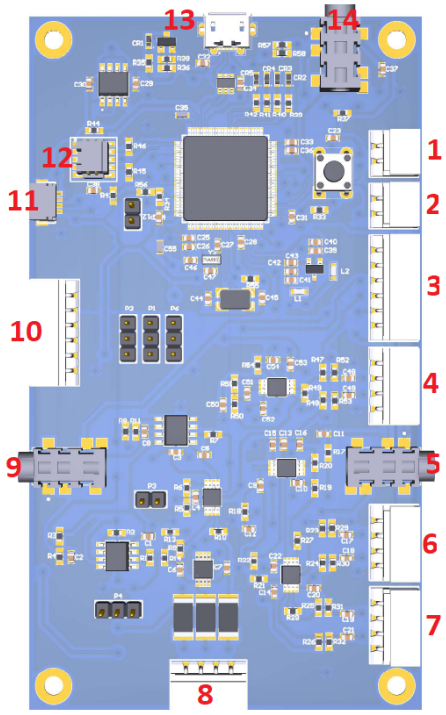


Figure 4.12: Speech and ECG Measurement System

1. Digital analog converter output. It is used to construct an analog ECG signal which passes through digital filters.
2. Digital analog converter output. It is used to construct analog speech signal which passes through digital filters.
3. It will be used for future works. It will be used if the analog audio signal is needed.
4. It is used for noise reduction. The omnidirectional microphone will be connected to this port. The device will convert it to digital and use it on an adaptive filter algorithm.
5. Analog Microphone Array output. It can be used to take speech signals through the audio jack of the PC.
6. Directional microphone array first input
7. Directional microphone array second input
8. It is used to attach copper plate for Lead-I ECG measurement.
9. Analog ECG Signal output. It can be used to take ECG signals through the audio jack of the PC.
10. It can be used if the system wants to feed with a battery to achieve low noise recordings. A jumper is attached to P2, P1, and P6 to configure the system to work with battery mode. These jumpers separate the regulator voltages powered by USB with battery voltages.
11. It is used for future work. It will be used for fingerprint data collection.
12. JTAG input for the microcontroller to debug and program it.
13. USB output. Speech and ECG signals are converted to ASCII format and sent to PC. ECG sampling rate is 1000 Hz, whereas the audio sampling rate is 24 kHz.

14. It will be used for future work. When the audio jack is connected, it will automatically detect the jack and will convert the analog signal to digital.

4.3 ECG Signal Recording Results

Raw ECG signal, ECG signals applied various filters, and frequency spectrum is shown in the following figures 4.13, and 4.14.

In the figure 4.13, a 12-bit Raw ECG signal is taken by the proposed speech and ECG Measurement Device. The signal is shown in the first 6 seconds of the record, and the final 2 seconds of the record are caused by not holding the copper plates. While constructing the database, this type of undesirable signal must be eliminated. Some traditional devices offer a lead-off detection circuit when the electrodes do not touch the body. In our proposed system, an ECG spikes and inconsistent beats detection algorithm is constructed to eliminate inconsistency of data instead of using a lead-off detection circuit. This algorithm will be explained later.

In figures 4.14, filter operation and its effect on the FFT are shown step by step. In figure 4.14(a) and 4.14(b), the result shows us that there is a DC component over the ECG signal. DC component's magnitude is both seen in the frequency spectrum of the signal or signal in the time domain as 1000. In figure 4.14(c) and 4.14(d), 0.5 Hz 6^{th} order high pass Butterworth filter is applied onto the raw signal, and it is seen that the DC component is removed and the signal is brought in the zero levels. When the scale of the signal changes due to removing of the DC component, now it can be easily seen that there are 50 Hz and its harmonics over the signal. To remove 50 Hz and 100 Hz, Notch filters are applied onto the signal, and the result is shown in the figure 4.14(e), 4.14(f). The 150 Hz and higher harmonics are eliminated after the window-based FIR low pass filter is applied onto the signal, and the result is shown in the figure 4.14(g), 4.14(i). And finally, a smoothing operation is applied, and the result is shown in the figure 4.14(h) and 4.14(j).

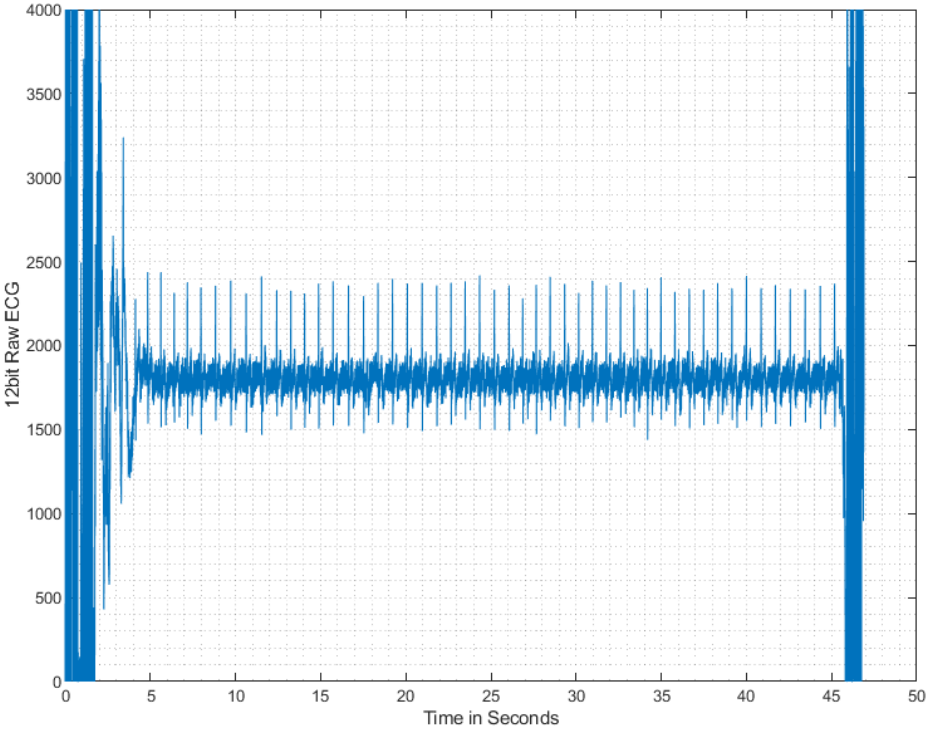
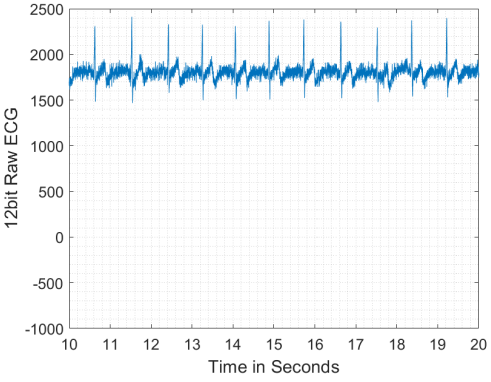
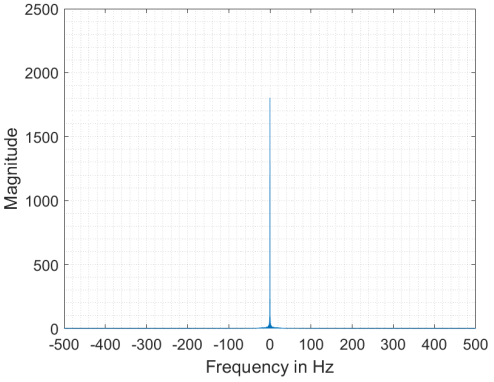


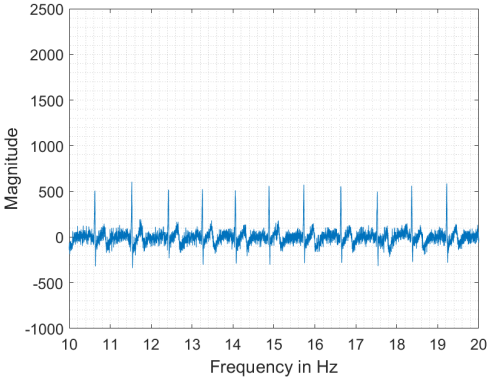
Figure 4.13: 12-bits Raw ECG Signal



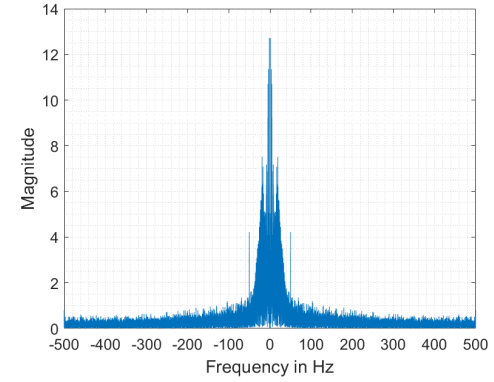
(a) Raw ECG Signal



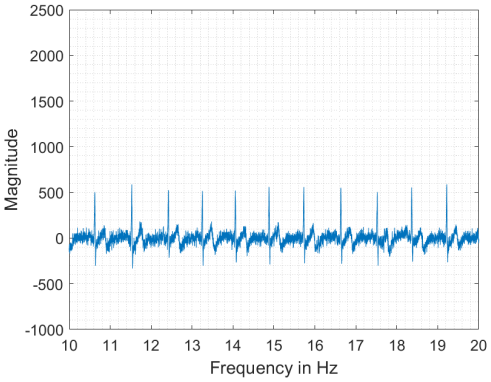
(b) DFT of Raw ECG signal



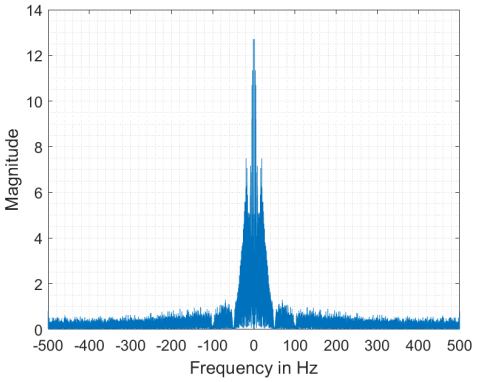
(c) Filtered ECG Signal (High pass)



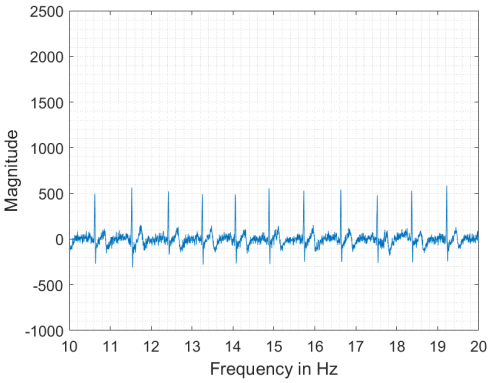
(d) DFT of Filtered ECG Signal (High pass)



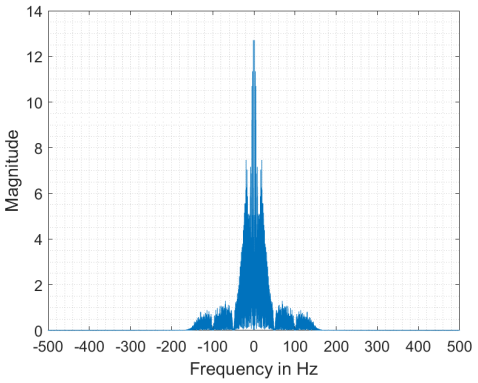
(e) Filtered ECG Signal (Notch)



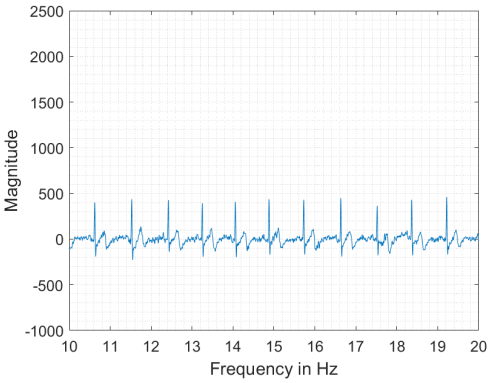
(f) DFT of Filtered ECG Signal (Notch)



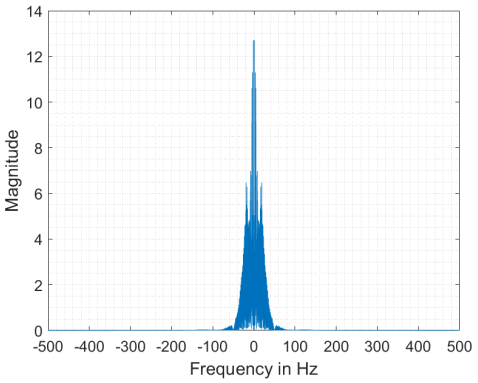
(g) Filtered ECG Signal (Low pass)



(h) DFT of Filtered ECG Signal (Low pass)



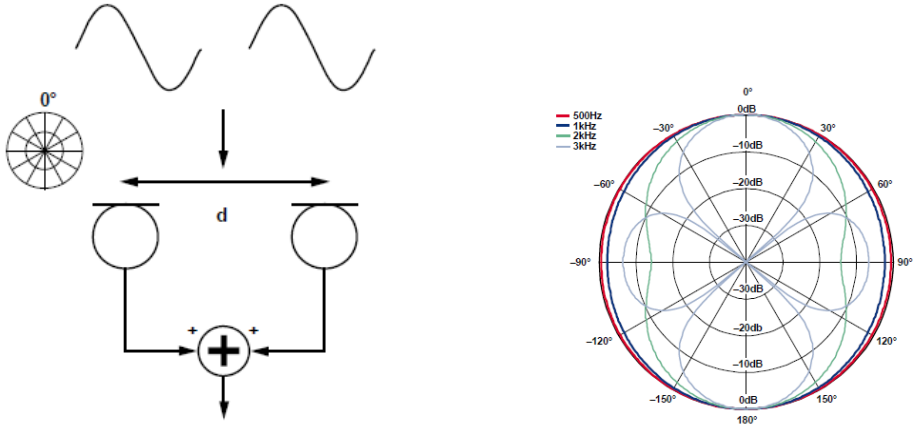
(i) Filtered ECG Signal (Smoothing)



(j) DFT of Filtered ECG Signal (Smoothing)

Figure 4.14: ECG Filtering Processes

4.4 Design of Microphone Array Beam



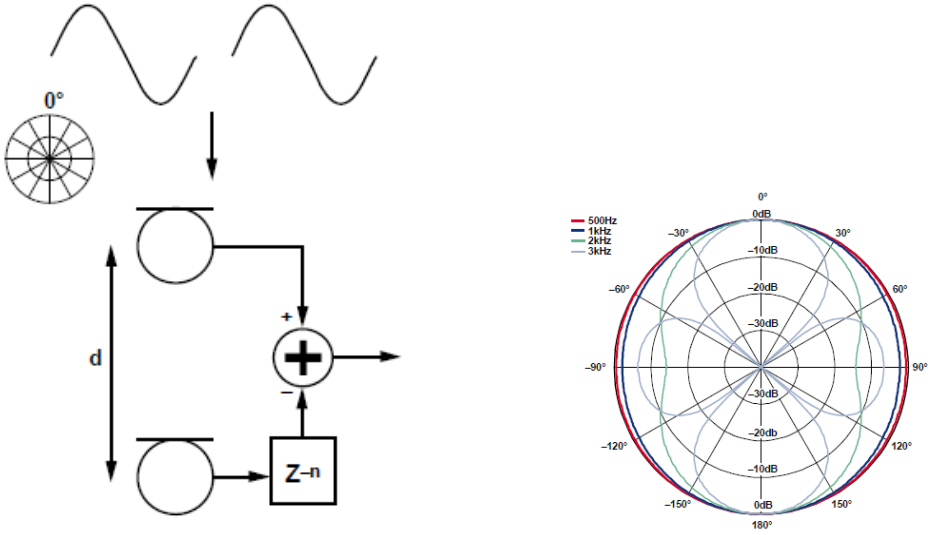
(a) Broad Side Array Topology (2-Microphone)

(b) Response of a 2-Microphone Broadside Array with 75 mm Spacing

Figure 4.15: Design of Broad Side Microphone Array Beam [50]

A broad-side microphone topology is a type of microphone array in which the microphones are lined in placed perpendicular to the preferred direction of sound waves (see in figure 4.15(a)). In that figure, d represents the dimension, and it corresponds to the spacing between the two elements of the array. The sound from the broad-side of the array is what is usually desired to be picked up [50].

The basic process of broad-side array topology is that summing the signal recorded from microphones in the array. The disadvantage of such array topology is that it can only attenuate the sound coming from the side of the array. The rear-facing response always matches the front response since there is nothing differentiating pressure waves approaching the microphones from the front and the rear because of the asymmetry of the array. A broad-side array topology is useful in applications where there is not much sound incident from behind or above and below the array, such as for television mounted on a wall [50].



(a) Endfire Array Topology (2- Microphone)

(b) Response of a 2-Microphone Endfire Cardioid Beamform

Figure 4.16: Design of End Fire Microphone Array Beam [50]

Endfire array topology consists of multiple microphones arranged in the line and put in the direction of desired sound. When the front microphone in the array where the first sound reach is summed with an inverted and delayed signal from the rear microphones, this configuration is called a differential array. Figure 4.16(a) shows a 2-microphone end-fire differential array with a distance (d) between two microphones and the rear microphone delayed by n samples before the subtraction and inverting block. This topology can be used to create cardioid, hyper-cardioid, or super-cardioid pickup patterns, where the sound from the rear of the array is greatly attenuated [50].

The sound picked up by the different microphones in the end-fire array differs only in the arrival time, assuming far-field propagation that can be approximated by a plane wave. For creating a cardioid pickup pattern, the signal from the rear microphones should be delayed by the time that it takes the sound waves to travel between the two microphone elements. It gives the system designer two degrees of freedom in designing an end-fire beamform: the distance between the microphones and the delay applied in the processor. In many audio applications, the choice of minimum delay time totally depends on the sampling rate of the

signal pick up from microphones. When the signal is quantized by 48 kHz, then the minimum delay is found as 21 μs for each sample. At 20 °C, the speed of sound in air is 343 m/sec, so sound wave travels about 7 mm in 21 μs [50]. Designing parameters and equation is given as

Distance of Sound Travels in a Specified Time:

$$d = c \times t \tag{4.1}$$

Microphone Spacing to Match an n-Sample Delay:

$$d = n \times c/Fs \tag{4.2}$$

Time Delay for an n-Sample Delay:

$$tD = n/Fs \tag{4.3}$$

where

c: it represents the speed of sound in air, in m/sec (343 m/sec at 20 °C)

d: it represents the distance in meters

t: it represents the time in seconds

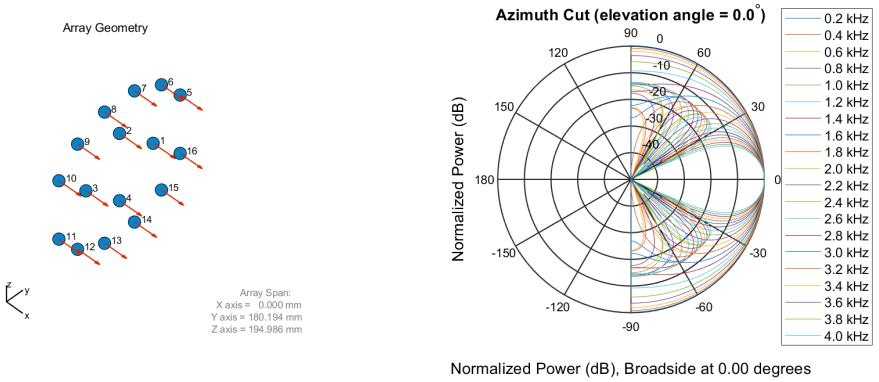
n: it represents the number of samples of delay in DSP.

Fs: it represents the sampling frequency in hertz

tD: it represents the time delay in seconds

Design specifics and experimental work for broad-side microphone array topology are given in figure 4.17(a). Placements of 16 Microphones in broad-side topology are designed by using Matlab. 4 microphones are placed in the center of the plate in the form of polar coordinate is given radius as 0.050 meters, Angles for each 0, 90, 180, 270. 12 microphones are placed with a radius of 0.1 meters to the center,

and angles are 25, 51, 77, 102, 128, 154, 205, 231, 257, 282, 308, 334, respectively. In the proposed method, this broad-side microphone array placement is used.

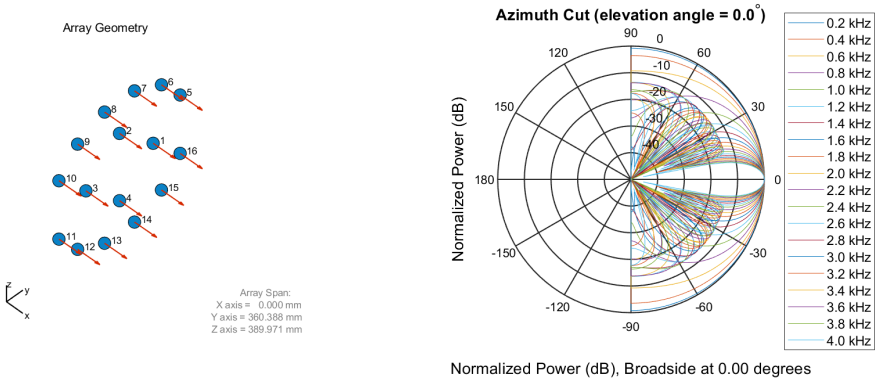


(a) Broad-Side 16 Microphone Array Beam Placement in 3D Space

(b) Response of 16- Broad side Microphone array beam which replaced specific order

Figure 4.17: Design of Broad Side Array Beam with 16 Microphone

In figure 4.17(b), frequency response of Broad-side 16 Microphones is shown. The lower frequency component of audio signals, which are vowels, pass through in every angle and direction. However, the higher the frequency components of audio, the higher the exposure to angular and directional suppression. If the microphone distances in the array increase, the microphone array's selectivity to the frontal voice source also increases. In figure 4.18(a), 4.18(b), Frequency spectrum of microphone array which designs in different radius distances are shown.



(a) Broad-Side 16 Microphone Array Beam Placement in 3D Space

(b) Response of 16- Broad side Microphone array beam

Figure 4.18: Design of Broad Side Array Beam with 16 Microphone (Max Performance)

Specifications of maximum performance (maximum frequency component is assumed as 4khz):

R=0.1 m for 4 microphones and angles are same as previous application

R=0.2 m for 12 microphones and angles are the same as the previous application

4.5 Fingerprint Recording Results

(a) GT-521F32 Fingerprint Sensor (b) Grayscale image taken by fingerprint sensor (c) Binary image extracted by proposed algorithm

Figure 4.19: Fingerprint Measurement

GT-521F32 sensor is a TTL-level fingerprint sensor. (see in figure 4.19(a)) It produces 450 dpi resolution fingerprint image in the figure 4.19(b). The image extracted from the fingerprint sensor is the first store in the proposed system, then it is sent to PC by UART communication protocol. After that, the grayscale image is converted to a binary image to use in future work.

Chapter 5

Speech and Fingertip ECG Signal based Person Recognition System

5.1 Block diagram of Speech and Fingertip ECG Signal based Person Recognition System

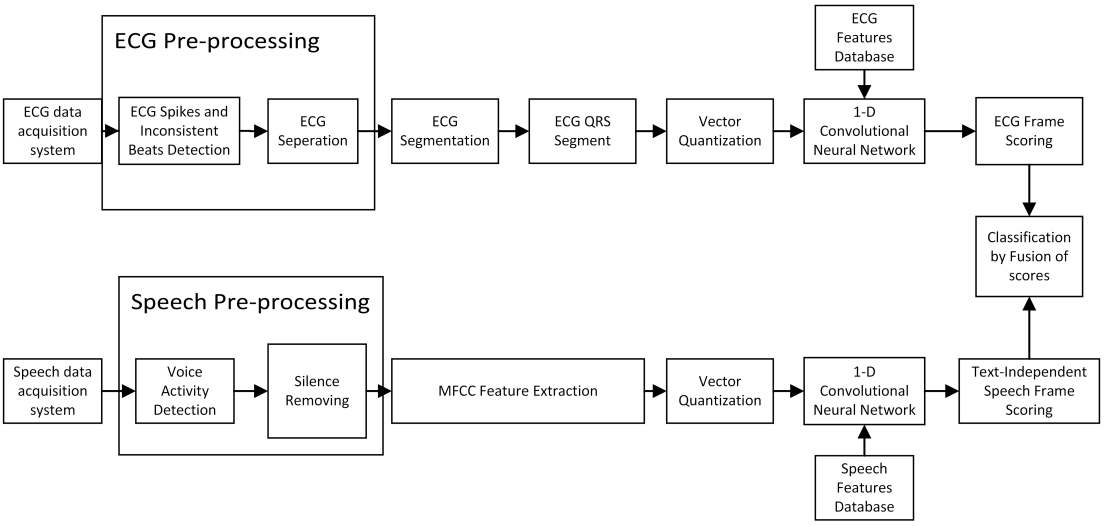


Figure 5.1: Block diagram of Speech and Fingertip ECG Signal based Person Recognition System

The algorithm in Figure 5.1 were designed to identify an individual based on his/her fingertip ECG and speech signal. The fingertip ECG signal and speech signal were acquired using a hybrid acquisition system that contains an instrumentation amplifier and an internal microphone. The fingertip ECG data of a person were collected by the patient's thumbs while speech signals were collected

using an internal microphone. The algorithm can identify a person with a speech signal which does not have to be a fixed phrase. It means that the algorithm can differentiate speakers with a text-independent speech signal. The following section only explains the person identification algorithm by using two different modalities and their fusion in detail. The decision rule of genuine people identification and imposter people rejection were explained in section 6.1 in details.

Fingertip ECG signals are measured from 58 individuals in different weeks by using the proposed measurement system. For each week, approximately 1 minute ECG signals were recorded, and each ECG signal was passed through a filtering process such as 0.5 Hz high pass filter, 50 and 100 Hz notch filter, 150 Hz notch filter, and smoothing filter. From now on, ECG signals measured in the first week will be used for the training phase, whereas ECG signals measured in the second week will be used for the testing phase. In the first part of training process, Fingertip ECG signals taken from 60 individuals are applied to the proposed system (see in figure 5.1). Then ECG Spikes and Inconsistent beats detection algorithm is applied to ECG signals to prevent unrelated noisy signals and uncorrelated ECG signals because of the movement of individuals. Some spikes may occur on the ECG signal, which causes some part of the ECG waves to corrupt because the person does not hold the Fingertip ECG connector steady. This block diagram finds the uncorrelated ECG signal, and if the total signal is not enough for the recognition system, it deletes the person from the database. If it is possible, the records will be taken again. However, in our proposed system, we continue with ECG signals of 54 individuals because the ECG signal of 4 people was eliminated in this block. After the ECG Spikes and Inconsistent beats detection, if the separated ECG signals are lower than 10 seconds, it will eliminate them to ensure the integrity or unity of ECG waves. After the ECG separation block, ECG segmentation is applied, and P waves, QRS waves, T waves, and their intervals Q, R, and S points, QT intervals, PR intervals, PR segments, ST segments were found. QRS Complex intervals were found as significant features for the ECG

recognition system. For this reason, the ECG signal of each individual was arranged into vectors placing each R point of the ECG wave into the fixed location. Fixed number of the left (165 samples) and right side (319 samples) of R points are clipped and defined that area as P-QRS-T wave. Then vector quantization is applied on the P-QRS-T wave of each individual to decrease the number of features. It captures the significant P-QRS-T wave of overall data and prevents the over-fitting problem in the machine learning algorithms. For each person, 16 significant P-QRS-T waves were found, and they were normalized in the range of 0 to 1. Then the system was trained by using a 1-D CNN algorithm.

In the second part of the training process, the RedDot speech database, which was released on August 17 in 2015, is used. The database was initiated with the collaboration of multiple sites during INTERSPEECH 2014. It was set out to collect speech signals throughout mobile crowd-sourcing with the benefit of a potentially wider population and greater diversity. It consists of 62 speakers, including 49 male speakers and 13 female speakers in 21 countries, and English was used as spoken language. Some of the signals in the database do not only contain speech of person but also contains background noises such as mouse-clicking sound, music, other people's conversation, microphone internal noises, whereas some signals contain pure voice. To further improve the system, this database which forces the system performance, was used. In the database, the speech file of each person was separated into folders, and each folder contains a speech signal where the given text is read by an individual. Some texts are the same for each speaker, whereas some texts are unique to individuals. Each text can be read approximately 3 seconds to 7 seconds, so each speech file contains approximately 5 seconds of speech data. The number of signals is also different for each person. The minimum number of speech files in the folders is found as 24, whereas the maximum number of speech files is found as 744. Approximately every folder contains 246 speech files by taking the mean of the number of speech signals for each folder. Then randomly, 54 folders were selected, and speech files in the folders separated into two parts where half of the speech files are

used for training whereas the other half is used for testing the system. Voice activity detection is used for detecting the voice and unvoiced (noisy) parts of the speech file. After that, the unvoiced part of the speech is separated into 20 ms frames, and the standard deviation of each frame which consists of the noisy signal, is found. We took the median of these standard deviations and used it as a threshold. In the speech and background noise block, speech signal separated into 20 ms frames with an overlap of 65%, and if the standard deviation of the speech frame is lower than the threshold, which multiplies by 2, the speech frame will be discarded. Then MFCC feature extraction method is applied to speech frames. After that, vector quantization is applied to speech frames to decrease the redundancy. For each person's every 10 seconds speech signal, 32 significant MFCC vectors were found, and they were normalized in the range of 0 to 1. Then the system is trained by using a 1-D CNN algorithm.

In the test phase, P-QRS-T waves of ECG signals recorded in the second week were found, and redundant P-QRS-T were eliminated. At the same time, MFCC features of speech signal were found, and redundant MFCCs were eliminated. After that, both speech and ECG features are normalized in the range of 0 to 1. Then both features are applied to the 1-D CNN classification method, and person id is found by comparing the scores, whereas the ratio for speech features were 3 and ECG features were 1.

In the following sections, blocks forming the proposed system will be explained in detail.

5.2 ECG Spikes and Inconsistent Beats Detection

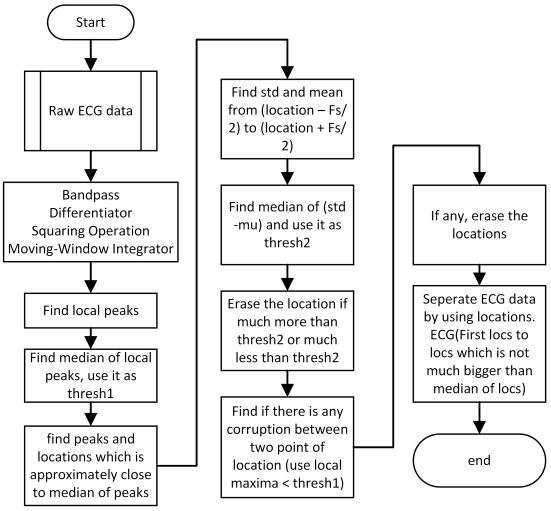
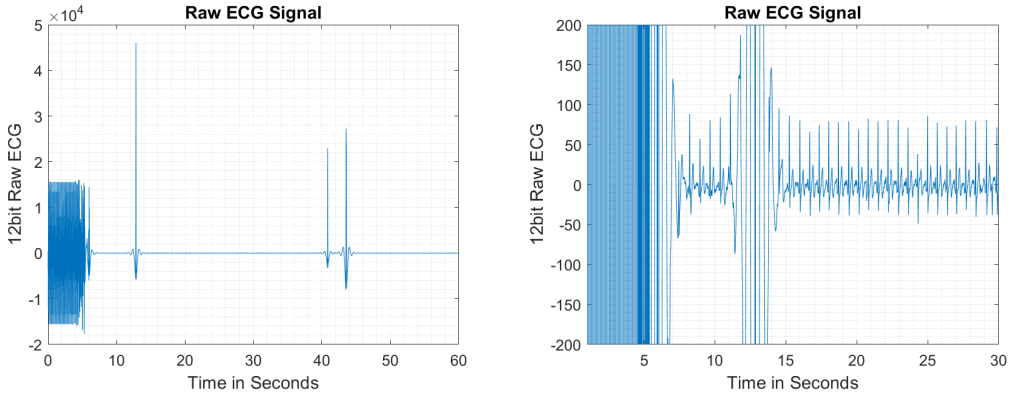


Figure 5.2: ECG Spikes and Inconsistent Beats Detection Block Diagram

ECG Spikes and Inconsistent beats detection algorithm block diagram is shown in the figure 5.2. This algorithm uses for eliminating inconsistent ECG signals or undesired signals. The algorithm derived for the inability of portable ECG measurement device that does not differentiate the signal when the copper plates are not held from ECG signal. Later on, the patient's movement is seen, which affects the shape of ECG waves changing. This kind of inconsistent data affects the performance of the recognition algorithm. For this reason, ECG Spikes and Inconsistent beats detection method is proposed. The algorithm works with the principle of finding and choosing the most repetitive peaks as R peaks. Then ECG signal is framed by using R peaks, and inconsistent ECG waves are eliminated because the standard deviation of an inconsistent signal is much lower or higher than the average standard deviation of ECG frames. In the end, the location of inconsistent ECG waves and undesired signals were deleted and were separated into sub vectors. The prerequisite for the algorithms to work efficiently, a total of consistent ECG signals in the record must be higher than undesired signals or inconsistent beats. If the requirement does not meet, the algorithm can't judge which signals are desired or not. ECG spikes and inconsistent beats detection algorithm is reviewed step by step in the below:

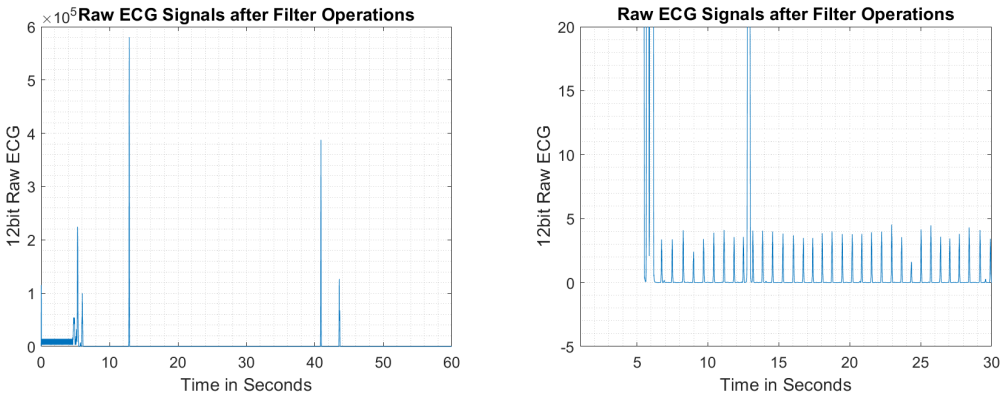
1. Firstly, Raw fingertip ECG signal, which is measured by the proposed Fingertip ECG acquisition system, is applied to the algorithm.



(a) Original Size (b) Zoom In

Figure 5.3: Raw ECG Signal

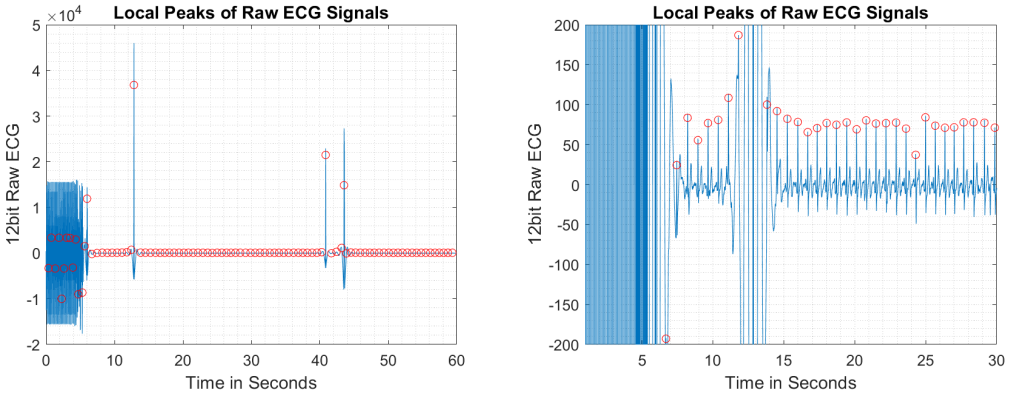
2. Then Raw ECG data pass through a Bandpass filter whose cut-off frequencies are between 5 Hz to 15 Hz because desirable QRS energy is approximately in these frequencies, and what we need to do is to find QRS peaks of ECG data and reject the other type of peaks. After that, we used a differentiator to provide QRS complex slope information, and it suppresses the low-frequency component of ECG data which are P and T waves. After that, we used squaring operation to make all the data point positive. Then, we used moving window integrator so that we sum of the N width data to make three peak points where the maximum peak point is QRS complex. If the width of the window integrator is too large, the output will merge QRS peak with P peak and T peak. If P wave, T wave, and QRS wave are merged, this will reduce the total amplitude of the peak because there is an isoelectric line between these peaks which reduces total peak amplitude. If the width of the window integrator is too small, it will cause additional peaks, which we do not want to. These four filters are found by PAN-TOMPKINS et al. [51] to find local ECG QRS peaks.



(a) Original Size (b) Zoom In

Figure 5.4: ECG signal after filter operations

3. Then local peaks are found for every window (N) sample by using the output of the Pan-Tompkins filters. Window length (N) is defined as 850 samples for 1 kHz sampled data. The number 850 is found by considering the normal heart rate of a person is 70 beats per second in a minute. The formula $(60seconds/70beats) \times Fs$ gives us the window length.



(a) Original Size (b) Zoom In

Figure 5.5: Local Peaks of the ECG signal

4. In the local peaks block, the most repetitive ones are the ECG QRS (R peaks); other noisy peaks are not stationary like ECG R peaks. Because of that, the median of local peaks, which are the most repetitive ones is found, and being saved as threshold 1.

5. After that, the other type of peaks and their location are erased if they are not in the range of ECG peaks.

6. From this on, most of the ECG peaks correctly selected, but it also contains ECG waves that were caused by individual movement. We can't expect patients to hold the proposed ECG device steady. For this reason, additional measures were taken. The best way to analyze the differences between consistent and inconsistent ECG waves is to find standard deviations. Firstly, ECG waves which include P wave, T wave, isoelectric lines, and QRS complex, were framed into vectors by considering the location of R peaks found from step 5. Then standard deviation and expectation of each vector were found.

7. Then, the median of standard deviations of ECG vectors were calculated by using the formula below.

$$\sigma_k = \sqrt{\sum_{i=1}^{N} \frac{(x_{ik} - \mu)^2}{N}} \tag{5.1}$$

where,

σ_k represents standard deviation of ECG frame

x_i is the values of single frame

μ is the mean of single frame

N is the number of ECG frame

Then the value of σ_k is sorted ascending. By using the equation, the median of standard deviation was found.

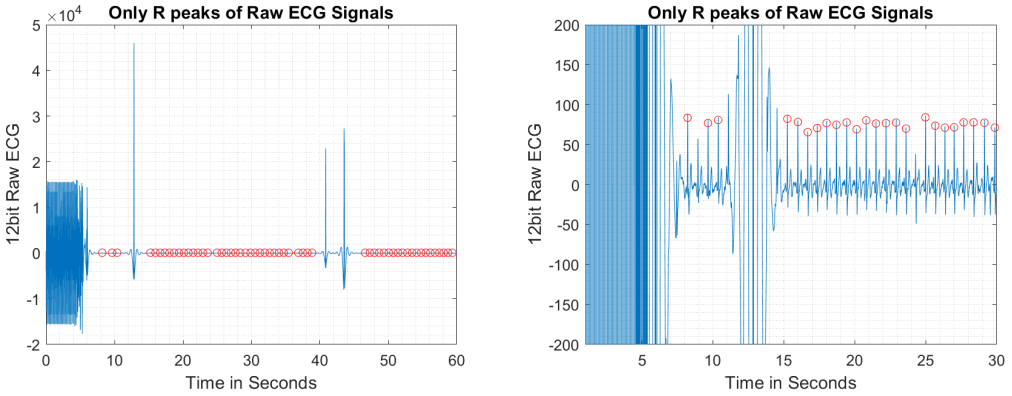
$$m = \sigma_{sort}\left[\left\{\frac{N+1}{2}\right\}^{th}\right] \tag{5.2}$$

Then m is saved as threshold 2.

8. In this section, QRS peaks, and their locations were deleted if their standard deviations are much more than threshold 2 or much less than threshold 2.

This will ensure deleting the corrupted ECG signals caused by the movement of fingers.

9. In some cases, some corrupted ECG data and noise are not removed. To make sure to remove them, local maxima for each QRS point were found, and if the local maxima points were more than threshold 1, these locations were erased.



(a) Original Size

(b) Zoom In

Figure 5.6: ECG R peaks

10. Finally, pure ECG signal points and their location were extracted. However, some of these points are not close to each other. We don't want to touch or eliminate the integrity (unity, wholeness) of ECG signals. Because of that, if the locations of ECG QRS peaks are much bigger than the nearest QRS points, we will choose it as the local last location. Then ECG signal puts into vector considering first peak location and local last peak location. After that new ECG vector's first location will be previously found local last location. This will continue until last ECG peak arranged into vectors (see in figure 5.7).

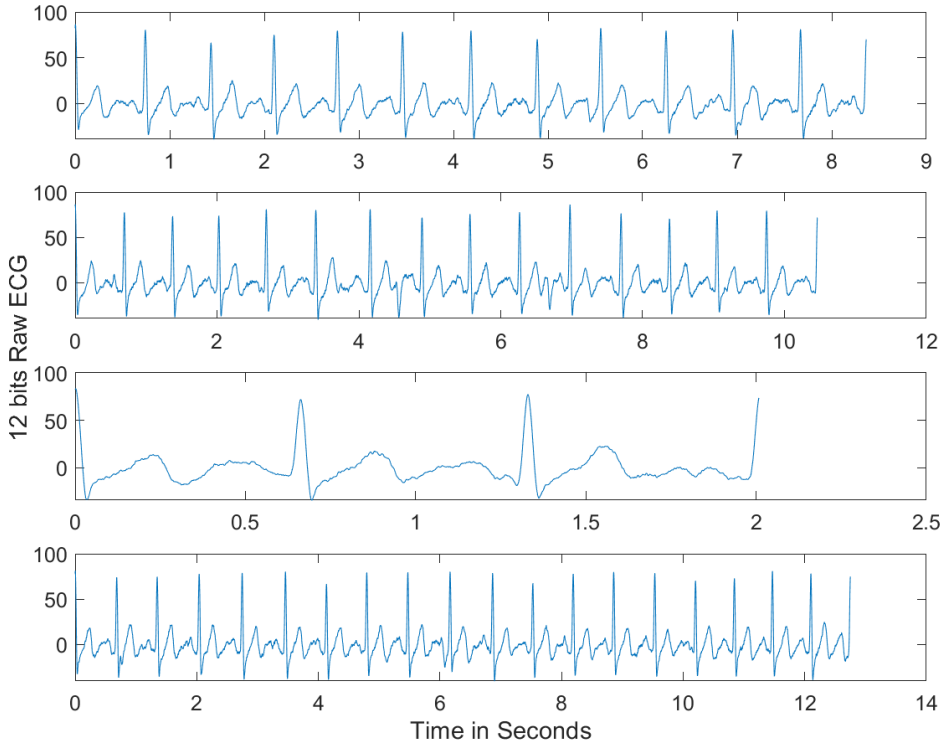


Figure 5.7: ECG Signal split into vectors

Pseudo-code for ECG Spikes and Inconsistent Beats Detection Algorithm is given in the below

Algorithm 1 ECG Spikes and Inconsistent Beats Detection Algorithm

Inputs:

Fingertip ECG files from the databases

Outputs:

Noise-free ECG signal vectors

Algorithm:

Apply bandpass, differentiator, squaring operation, and moving-window integrator to ECG signal, and obtain a signal which contains QRS information [51].
Define window length $W = 850$ where $F_s = 1000$
Frame the filtered signal into W_N vectors
for each vector i=1 to N **do**
    $M_i \leftarrow$ Find maximum value,
    $l_i \leftarrow$ Find index of maximum value,
    $l_i \leftarrow l_i -$ filters group delay
end for
Sort M into descending order
Find value of M in the middle, Store it into $Th1$
Define S_1 as $Th1/35$
for each vector i=1 to N **do**
    if $M_i < (Th1 - S_1)$ && $M_i > (Th1 + S_1)$ **then**
        Erase l_i from locations
    end if
end for
for ECG signal between l_i to l_{i+1} until $i = N - 1$ **do**
    if $l_{i+1} - l_i > 2F_s$ **then**
        continue
    end if
    Find $\sigma_i - \mu_i$ and Store it into V_i
end for
Sort V into descending order
Find value of V in the middle, Store it into $Th2$
Define S_2 as $Th2/5$
for ECG signal between l_i to l_{i+1} until $i = N - 1$ **do**
    if $l_{i+1} - l_i > 2F_s$ **then**
        continue
    else if $V_i < (Th2 - S_2)$ && $V_i > (Th2 + S_2)$ **then**
        Erase l_i from locations
    end if
end for
for ECG signal between l_i to l_{i+1} until $i = N - 1$ **do**
    if $l_{i+1} - l_i > 2F_s$ **then**
        continue
    else
        Find values of local maxima and Store them into L
    end if
    if $L_i < (Th1 - S_1)$ && $L_i > (Th1 + S_1)$ **then**
        Erase l_i from locations
    end if
end for
Separate ECG signal from location l_i to l_{i+1}, if not bigger than $2F_s$

From now on, the ECG vectors which are free from noise and corruption can be used on the proposed recognition method. In the ECG separation block, the duration of ECG signals in each vector is checked, and if any of the signals' duration is lower than 10 seconds, that signal vector will be eliminated from the database.

5.3 ECG Segmentation

ECG Segmentation algorithm block diagram is shown in the figure 5.8. This algorithm is used for extracting ECG fiducials features where will be later used for the ECG-based person recognition system. The algorithm finds QRS complex, P wave, T wave, QT-interval, PR-interval, respective points, and their locations, if possible PR-segment and ST-segment. In these fiducial points, the most significant feature is selected as a combination of P-QRS-T wave, and it will be used on ECG recognition system.

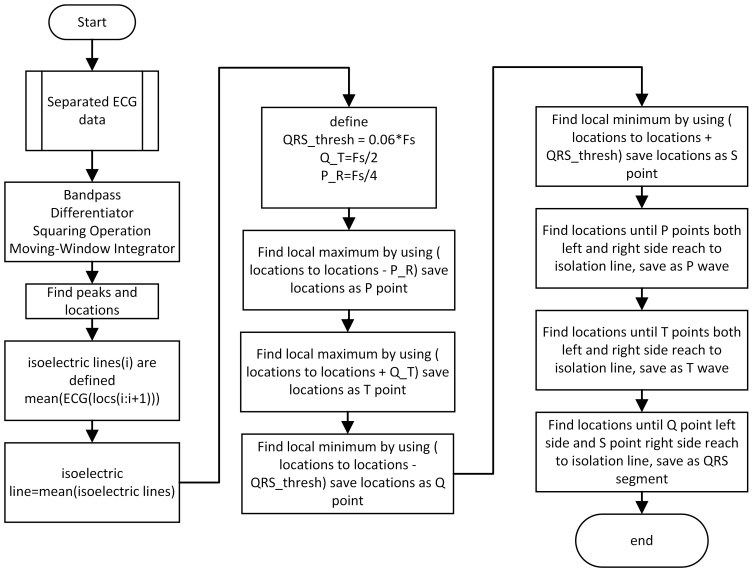


Figure 5.8: ECG Segmentation Block Diagram

ECG Segmentation algorithm is reviewed step by step in the below:

1. First noise-free, separated ECG data is applied into the algorithm.
2. This block is the same as the block in ECG Spikes and Inconsistent beats detection algorithm. The four filters are applied to the ECG signals so that QRS information can be picked out.
3. This block is also the same as the block in ECG Spikes and Inconsistent beats detection algorithm. The local peaks and their location by every 850 samples are found. However, this time, the result will give us the R point of the ECG signal.

4. After that, we find the ECG isoelectric line thresholds by taking the mean of every left and right side of QRS points.
5. Then, take the average of these thresholds and attain one threshold value for the isoelectric line.
6. QRS duration for a healthy person is defined as 0.06 seconds. So we select QRS_to thresh as 0.06 × ECG sample rate. This threshold will be used for finding Q and S points.

   Q_T threshold is defined as Fs/2, and it will be used for finding T peak.
   P_R threshold is defined as Fs/4, and it will be used for finding the P peak of the ECG signal.
7. Local maximum between starting from ECG R points to (ECG R points -P_R threshold) will give us P points.
8. Local maximum between starting from ECG R points to (ECG R points + Q_T threshold) will give us T points.
9. Local minimum between starting from ECG R points to (ECG R points – QRS_thresh) will give us Q points.
10. Local minimum between starting from ECG R points to (ECG R points + QRS_thresh) will give us S points.
11. We define P wave as starting from P point to left side of the isoelectric line to P point to right side of the isoelectric line.
12. We define T wave as starting from T point to left side of the isoelectric line to T point to right side of the isoelectric line.
13. We define the QRS complex as from Q point to left side of the isoelectric line to S point to the right side of the isoelectric line.
14. After finding the location of P waves, T waves, QRS complexes, P points, T points, R points, Q points, and S points, they are saved in the ECG structure.

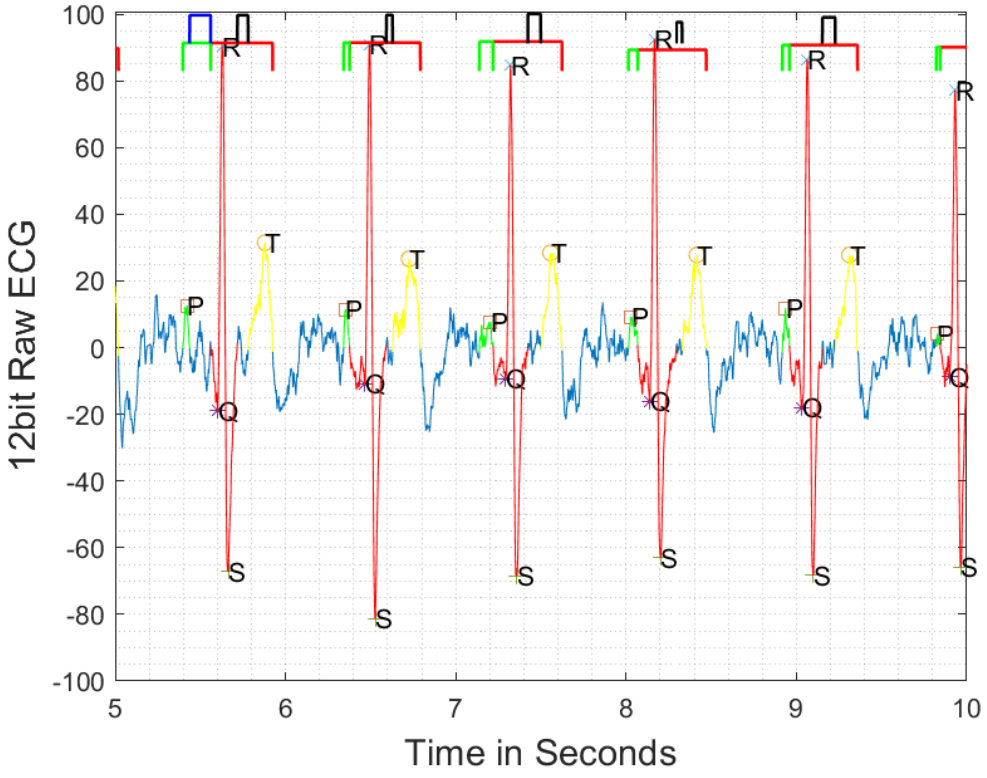
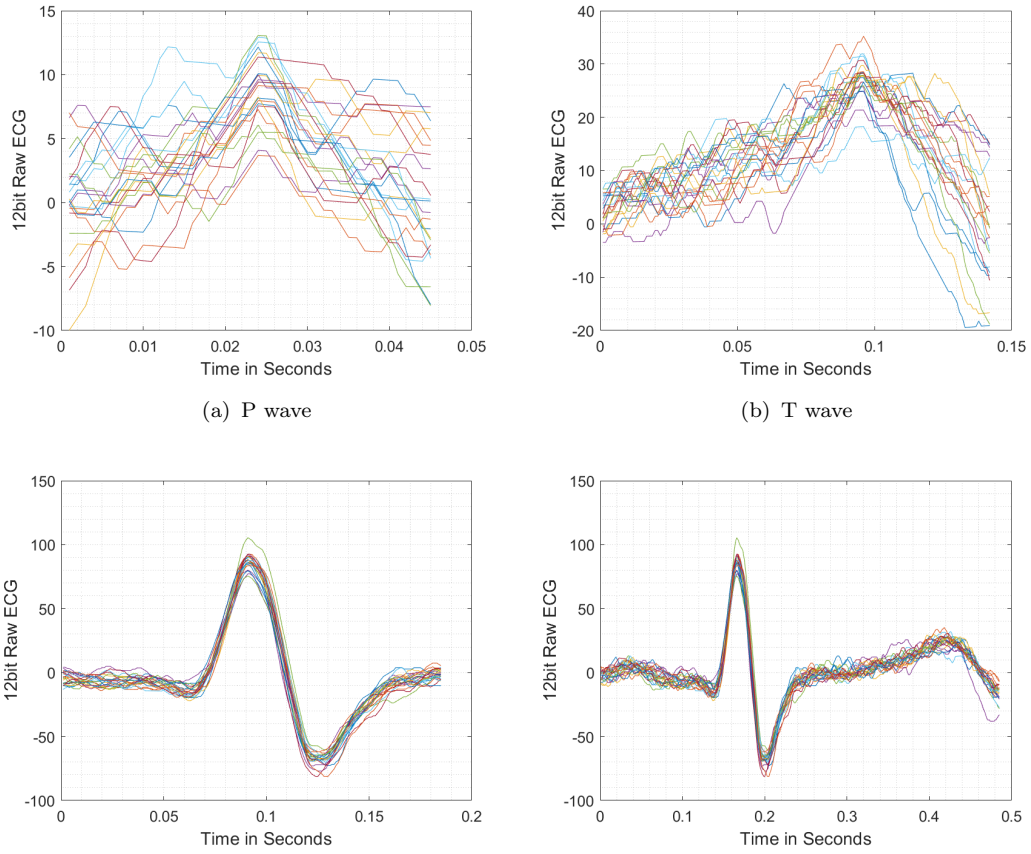


Figure 5.9: ECG segmentation

The fingertip ECG segmentation is shown in the figure 5.10 where P, Q, R, S, and T points were located and illustrated in different colors. The colors indicate the duration of P wave, QRS Complex, and T waves, whereas colorful lines on top of the ECG signal indicate the interval and segments. The green line indicates the PR interval, the red line indicates QT interval, the blue line indicates PR segments, and the black line indicates ST segments. Both PR segments and ST segments are hard to find in noisy ECG signals. Because of that algorithm, most of the time, can't find them. ECG waves are arranged in vectors by using related points, which are P points, T points, and R points, and these are shown in the following figures.



(a) P wave

(b) T wave

(c) QRS Complex

(d) P-QRS-T interval

Figure 5.10: ECG Segments

Pseudo-code for ECG Segmentation Algorithm is given in the below

Algorithm 2 ECG Segmentation Algorithm

Inputs:

Noise free ECG signals

Outputs:

Structure which contains ECG segments

Algorithm:

Apply bandpass, differentiator, squaring operation, and moving-window integrator to ECG signal, and obtain a signal which contains QRS information [51].
Define window length $W = 850$ where $F_s = 1000$
Define QRS_th $= 0.06F\_s$, QT_th $= F_s/2$, PR_th $= F_s/4$
Frame the filtered signal into W_N vectors
for each vector i=1 to W_N **do**
$l_i \leftarrow$ Find index of maximum value,
$\text{R\_lcs}_i = l_i -$ filters group delay
$\text{R\_pts}_i \leftarrow$ find value in the location of l_i
end for
for ECG signal between R_lcs_i to R_lcs_{i+1} until $i = N - 1$ **do**
$\text{P_pts}_i \leftarrow$max val between $(\text{R\_pts}_i - \text{PR\_th})$ to $(\text{R\_pts}_i - \text{QRS\_th})$
$\text{P\_lcs}_i \leftarrow$ index of max value P_pts_i
$\text{T_pts}_i \leftarrow$max val between $(\text{R\_pts}_i + \text{QRS\_th})$ to $(\text{R\_pts}_i + \text{QT\_th})$
$\text{T\_lcs}_i \leftarrow$ index of max value T_pts_i
$\text{Q_pts}_i \leftarrow$min val between $(\text{R\_pts}_i - \text{QRS\_th})$ to R_pts_i
$\text{Q\_lcs}_i \leftarrow$ index of min value Q_pts_i
$\text{S_pts}_i \leftarrow$min val between R_pts_i to $(\text{R\_pts}_i + \text{QRS\_th})$
$\text{S\_lcs}_i \leftarrow$ index of min value S_pts_i
$\text{isoline}_i \leftarrow$mean value between R_pts_i to R_pts_{i+1}
    for ECG signal between j=1 until $j = (\text{R\_lcs}_{i+1} - \text{R\_lcs}_i)$ **do**
$\text{Q_iso}_i \leftarrow$location of first value exceed of $\text{Q\_lcs}_i - j > \text{isoline}_i$
$\text{S_iso}_i \leftarrow$location of first value exceed of $\text{S\_lcs}_i + j > \text{isoline}_i$
$\text{P_iso_l}_i \leftarrow$location of first value exceed of $\text{P\_lcs}_i - j > \text{isoline}_i$
$\text{P_iso_h}_i \leftarrow$location of first value exceed of $\text{P\_lcs}_i + j < \text{isoline}_i$
$\text{T_iso_l}_i \leftarrow$location of first value exceed of $\text{T\_lcs}_i - j < \text{isoline}_i$
$\text{T_iso_h}_i \leftarrow$location of first value exceed of $\text{T\_lcs}_i + j > \text{isoline}_i$
    end for
$\text{P_wave}_i \leftarrow$values between P_iso_l_i to P_iso_h_i
$\text{QRS_cmplx}_i \leftarrow$values between Q_iso_i to S_iso_i
$\text{T_wave}_i \leftarrow$values between T_iso_l_i to T_iso_h_i
$\text{QT_interval}_i \leftarrow$values between Q_iso_i to T_iso_h_i
$\text{PR_interval}_i \leftarrow$values between P_iso_l_i to Q_iso_i
$\text{PR_segment}_i \leftarrow$values between P_iso_h_i to Q_iso_i
$\text{ST_segment}_i \leftarrow$values between S_iso_i to T_iso_l_i
$\text{P-QRS-T_fixed}_i \leftarrow$values between $(\text{R\_pts}_i - 165)$ to $(\text{R\_pts}_i + 319)$
end for

5.4 Vector Quantization

5.4.1 K-mean Clustering

In scientific research, clustering analysis plays an important role. K-means is a widely used partition method in clustering [52]. It is a method of vector quantization that aims to "n" observation into "k" clusters in which every one of the observations belongs to a cluster with the nearest mean (cluster centroid, or cluster center). K-mean clustering minimizes within-cluster variances by using squared Euclidean distances (see in figure 5.11). The objective of k-mean

clustering is given as

$$arg\,min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\| = arg\,min_S \sum_{i=1}^{k} |S_i|\, Var S_i \tag{5.3}$$

,given set of observations $(x_1, x_2, x_3, \ldots, x_n)$, where each observation is a "d" dimension real vector, n is the number of observation ,k is the number of cluster, $_i$ is the mean of points in S_i

The standard k-means algorithm uses the iterative refinement technique. The algorithm has two steps [53]:

First step: Assign each observation to the cluster

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\} \tag{5.4}$$

,where given initial set of k means $(m_1^{(1)}, \ldots, m_k^{(1)})$

,each x_p is assigned to $S^{(t)}$, even if it assigns into one or more clusters

Second (update) step: Recalculate means or centroids for observations that were assigned to each cluster

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{5.5}$$

The algorithm has converged when there is no longer a change in the assignments. The algorithm doesn't guarantee to find the optimum. It only assigns the objects to the nearest cluster by distances.

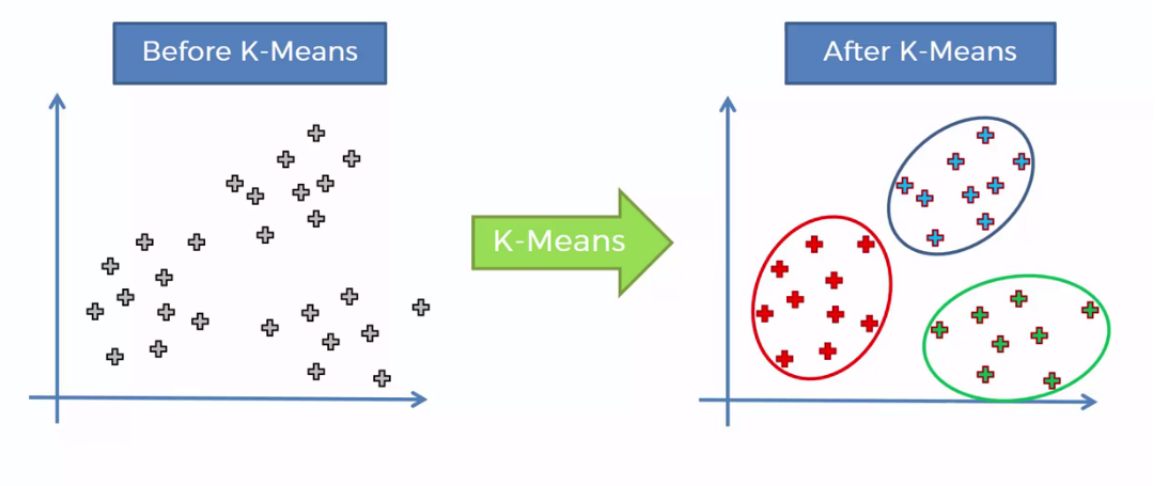


Figure 5.11: K-mean clustering

5.4.2 K-means Vector Quantization Algorithm

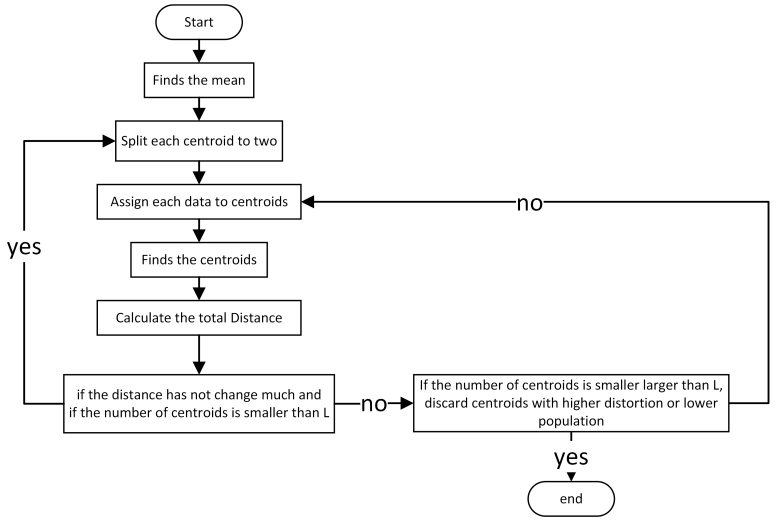


Figure 5.12: Block diagram of K-means Vector Quantization (Esfandiar Zavarehei, 2006)

The algorithm was developed by Esfandiar Zavarehei at Brunel University in 2006, and it utilized the following steps.

The algorithm takes m x n column-based matrix and L, which is the maximum number of centroids needed to find in the m x n matrix.

1. It finds the means of each column-based vector in matrices. For m x n matrix, it will find 1 x N central mean

2. Then it splits into two centroids by using the following formula.

   (central mean $-0.1\times$ central mean), (central mean $+0.1\times$ central mean)

3. Then, it assigns each of the data in the column-based matrix to that two centroids by using the squared Euclidean distance formula. The data assigns to the nearest centroid.

4. It finds new centroid using the data assigned to that centroid by taking the mean of the data

5. It calculates the total distances of each centroid by using the data assigned on that centroid. Then it calculates the sum of total distances for each centroid.

6. (Total distance – previous total distance)/previous total distance <threshold, where the threshold is starting with the value of 0.005, and this threshold value is decreased with (threshold $\times 0.75$) when centroids split into two.

   Because for each splitting, centroids will converge to specific locations, and improvement will need to increase for each splitting.

   If the new centroids were found and data were assigned to newly found centroids; however, if there is still no change in total distance, the number of centroids then split into four by using the following formula:

   (first centroid mean $-0.1\times$ first centroid mean), (first centroid mean $+0.1\times$ first centroid mean)

   (second centroid mean $-0.1\times$ second centroid mean), (second centroid $+0.1\times$ second centroid mean)

   If the total distance has an improvement when compared with the previous total distance, then it will continue to change the centroid by using the assigned data to that centroids. If it is not improving, split centroids into two, again.

7. Whenever the maximum number of centroids (L) are achieved, and there is no change in the total distance. Get to the next step.

In the next step, look at the value of the maximum number of centroids (L). It must be the power of 2. If the maximum number of centroids (L) is not the power of 2 then discards the highest distortion centroids or lowest population centroids. Until achieving the given L value.

8. After that, the algorithm finds the data close to each centroid and finishes its job.

5.5 Voice Activity Detection

Voice Activity Detection is a technique that detects the presence or absence of speech. It has often been applied in speech-controlled applications, which are operated by using speech commands. Most voice-based home automation systems are based on a VAD algorithm to detect speech and operate their task [54]. In the proposed method, we used the VAD algorithm, which was developed by Jongseo Sohn et al. [55]. The VAD algorithm assumes that the statistics of a background noise are stationary over a long period of time than the presence of a speech signal. They suggested a decision system which detects presence or absence of a speech by observing estimated noise statistics in current frame. Their decision rule derives from the likelihood ratio test (LRT) which estimates unknown parameters using the maximum likelihood (ML) criterion. In addition to their decision system, they proposed a hang-over scheme to minimize miss detections at weak speeches [55].

By the assumption that speech is degraded by uncorrelated additive noise, the decision rules have two hypotheses for a VAD to consider for each frame.

$$H_o : speech\ absent :\ X = N \tag{5.6}$$

$$H_1 : speech\ present :\ X = N + S \tag{5.7}$$

,where S, N, and X are "L" dimensional DFT coefficient vectors of speech, noise, and noisy speech

They adopted a Gaussian statistical model in which the DFT coefficient of each process is asymptotically independent Gaussian random variables. The probability density function conditioned on H_0 and H_1 are given as

$$p(X \mid H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} e^{-\frac{|X_k|^2}{\lambda_N(k)}} \tag{5.8}$$

$$p(X \mid H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi \left[\lambda_N(k) + \lambda_S(k)\right]} e^{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}} \tag{5.9}$$

,where k is the element of vector an $\lambda_N(k)$ and $\lambda_S(k)$ denote the variances of N_K and N_S, respectively.

The likelihood ratio for k^{th} frequency band is given as

$$\Lambda_k \triangleq \frac{p(X_k \mid H_1)}{p(X_k \mid H_0)} = \frac{1}{1+\xi_k} e^{-\frac{\gamma_k \xi_k}{1+\xi_k}} \tag{5.10}$$

,where ξ_k and γ_k called a priori and a posteriori signal to noise ratios (SNRs), respectively. They define as $\xi_k \triangleq \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma \triangleq \frac{|X_k|^2}{\lambda_N(k)}$

Then the decision rule is established from the geometric mean of the likelihood ratios for the individual frequency bands and given as

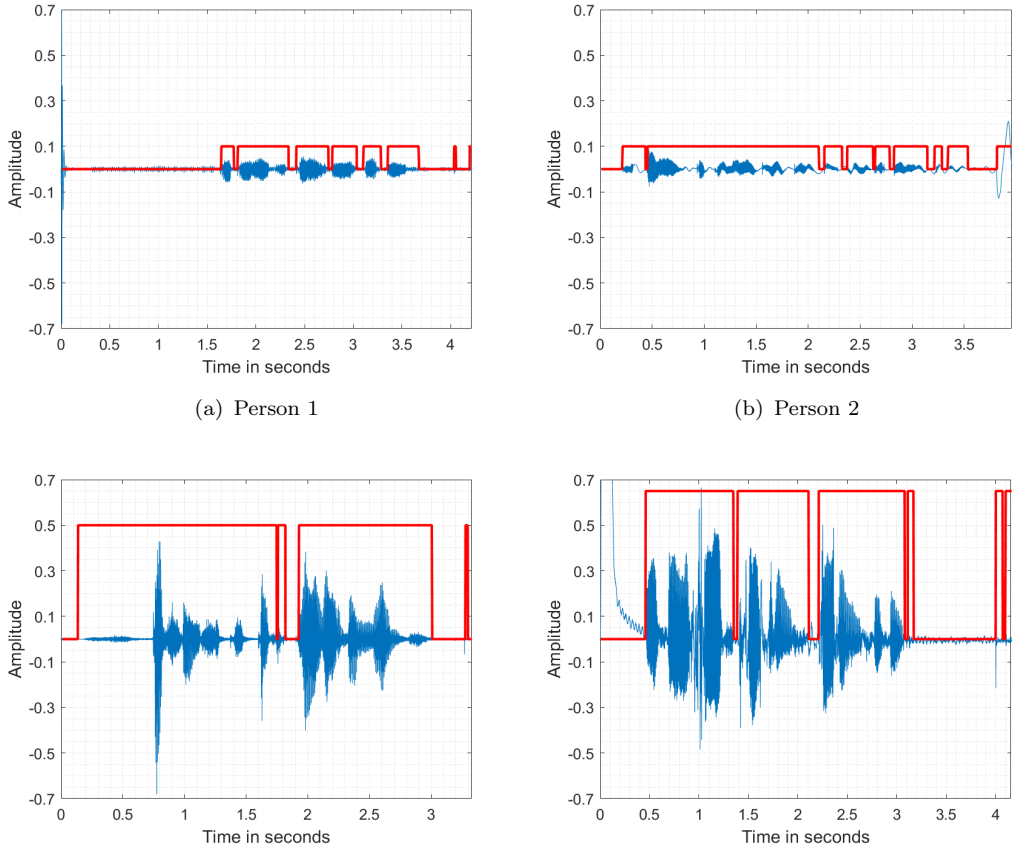
$$\log \Lambda^{(ML)} = \frac{1}{L} \sum_{k=0}^{L-1} \{\gamma - \log \gamma_k - 1\} > \eta \quad , H_1 \tag{5.11}$$

$$\log \Lambda^{(ML)} = \frac{1}{L} \sum_{k=0}^{L-1} \{\gamma - \log \gamma_k - 1\} < \eta \quad , H_0 \tag{5.12}$$

Left-hand side of the 5.11 and 5.12 can't be smaller than zero, for this reason likelihood ratio is biased to H_1. To reduce this biasing, a decision-directed a priori SNR estimation method is applied, and ξ_k becomes

$$\xi_k(n)^{DD} = \alpha \frac{\hat{A}_k^{\,2}(n-1)}{\lambda_N(k, n-1)} + (1+\alpha) P[\gamma_k(n) - 1] \tag{5.13}$$

,where n is the frame index, and $P[x] = x$ if $x \geq 0$, and $P[x] = 0$ otherwise, and $\hat{A}_k(n-1)$'s are the signal estimate amplitude of previous frame. This method provides smooth estimation of a priori SNR which reduces the fluctuation of the estimated likelihood ratio during noise periods.



(a) Person 1

(b) Person 2

(c) Person 3

(d) Person 4

Figure 5.13: Result of VAD algorithm

5.6 Mel Frequency Cepstral Coefficients (MFCC)

Mel-frequency Cepstral Coefficient (MFCC) is a method which is commonly used in automatic speech and speaker recognition system. It was introduced in the 1980s by Davis and Mermelstein [56] and has been used as modeling the behavior of the human auditory systems ever since. The Mel frequency cepstrum is a representation of the short-term power spectrum of a sound, based on linear cosine transform (DCT) of a log power spectrum on a non-linear Mel frequency scale. MFCCs are the coefficients that collectively make up a Mel Frequency Cepstrum [57]. Algorithm steps can be seen in Figure 5.14.

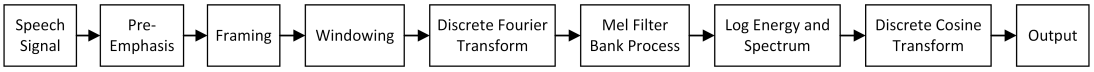


Figure 5.14: MFCC Block Diagram

Pre-Emphasis: It is a step that voice samples pass through a filter which emphasizes higher frequencies. It increases the energy of high frequency component of the voice. It is achieved by applying following equation onto the sampled voice:

$$y[n] = x[n] - \alpha x[n-1] \tag{5.14}$$

,where $0.9 < \alpha < 1$

In the proposed method, α is given as 0.97

Framing: Speech signals are non-stationary, and their temporal characteristics (such as the energy, zero-crossing rate, etc.) change very fast. So, if speech signals are taken as small frames, we can make an assumption that the speech signal will be stationary, and its characteristics within each frame will not vary much. In addition that, short-shifting between frames is essential for tracking continuity in the speech and not missing out on any abrupt changes at the end of frames. The process of shifting is called "overlap" or "step size". The voice signals are usually

segmented into frames of 20 ms ~ 40 ms with an overlap of 1/3 ~ 1/2 of its frame size.

In the proposed method, Speech signals are segmented into frames of 20 ms frames with an overlap of 65% its frame size.

Windowing (Hamming): It is used to minimize the spectral distortion caused by the transition of frames. It smooths the sharp frame transitions at both beginning of the frame and the end of the frame. Each frame has to be multiplied with a hamming window whose equation is given below.

$$w\left(n,\alpha\right)=\left(1-\alpha\right)-\alpha\cos\left(\frac{2\pi n}{N-1}\right),\ \ 0\leq n\leq N-1] \tag{5.15}$$

, where N represents the number of samples in the voice frame. The different α value corresponds to having different curves. In the proposed method, α is chosen as 0.46

Discrete Fourier Transform: It is used to convert voice frames from the time domain to the frequency domain. It is applied to voice frames which multiplied with hamming window previously. The equation is given as

$$X\left(k\right)=\sum_{n=0}^{N-1}x\left(n\right)e^{-\frac{j2\pi}{N}kn},\ \ 0\leq k<N \tag{5.16}$$

, where n represents the current voice sample, k represents frequency points which are divided into a length of N, N represents the maximum number of frequency points.

In the proposed method, the length of frequency point (N) is 256 points, where the frame length is 20 ms.

Mel Filter Bank Processing: It applies to find Mel-frequency scaling, which is a perceptual scale that helps simulate the working of human ears. It corresponds to better resolution at low frequency whereas less at high frequency. In

other words, the critical band's bandwidth that influences human ears to change between the frequencies, where it is linear below 1 kHz, and logarithmic above. By using a triangular filter bank, the distribution of these critical bands becomes linear. The position of these triangular filters is equally spaced along with the Mel frequency, which is related to the common frequency "f", and the following equation shows that relation.

$$mel\left(f\right) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{5.17}$$

Mel-frequency filter bank that includes M numbers triangular structure is defined as [58]

$$H_m\left(k\right) = \begin{cases} 0 & k < f\left(m-1\right) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f\left(m-1\right) \leq k \leq f\left(m\right) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f\left(m\right) \leq k \leq f\left(m+1\right) \\ 0 & k > f\left(m+1\right) \end{cases}, m = 1, 2, \cdots, M \tag{5.18}$$

, where M is the number of filters f is the uniformly spaced boundary points in Mel scale, and it is calculated with the following equation

$$f\left(m\right) = \left(\frac{N}{F_s}\right) {f_{mel}}^{-1}\left(f_{mel(low)} + m\frac{f_{mel(high)} - f_{mel(low)}}{M+1}\right),\ m = 0, 1, 2, \cdots, M \tag{5.19}$$

,where N is the size of DFT, F_s is the the sampling frequency and $f_{mel(low)}$, $f_{mel(high)}$ and f_{mel}^{-1} are defined as

$$
\begin{aligned}
f_{mel(low)} &= 1125 \ln\left(1 + \frac{f_{low}}{700}\right) \\
f_{mel(high)} &= 1125 \ln\left(1 + \frac{f_{high}}{700}\right) \\
f_{mel}^{-1}(f) &= 700\left(e^{\frac{f}{1125}} - 1\right)
\end{aligned}
\tag{5.20}
$$

,where f_{low} represents lowest, f_{high} represents highest frequency of the filter bank in Hz

In the proposed method, the frequency range of the triangular filter bank is chosen between 300 Hz to 3700 Hz, and the number of triangular filters is given as 20. The triangular filter bank is shown in the figure 5.15.

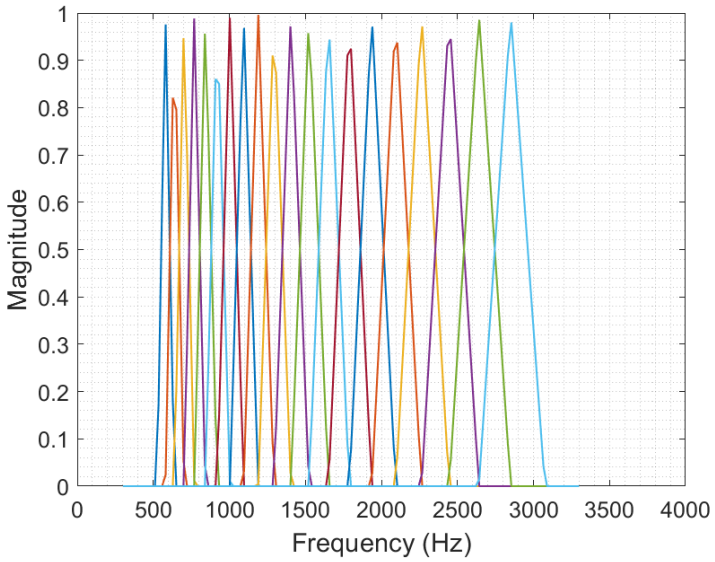


Figure 5.15: Triangular Filter Bank

Log Energy of Output of Filter Bank: The Mel-scale is achieved by multiplying each triangular filter with voice frames in the frequency domain. Then log energy is calculated by using the following equation.

$$
S(m) = \ln\left[\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right], \; 0 < m \leq M \tag{5.21}
$$

,where $|X_a(k)|$ is the magnitude of Discrete Fourier transform of voice frame and H_m is the triangular filter bank. $S(m)$ represents to Mel Spectrum coefficients.

Discrete Cosine Transform: MFCC features which represented with c(n) are found by applying Discrete Cosine Transform (DCT) on the Mel-Spectrum coefficients which were obtained from the previous process.

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi n (2m-1)}{2M}\right), \ 0 \leq n < M \tag{5.22}$$

, where M is the total number of MFCC coefficients.

After the DCT, the Cep Lifter operation is applied to reduce the undesirable components. In the following equation, the sinusoidal cep lifter is described.

$$w_i = 1 + \frac{L}{2} \sin\frac{\pi i}{L}, \ 0 \leq i < M \tag{5.23}$$

, where M is the total number of MFCC coefficients.

In the end, Liftered Ceptsral Coefficients found by multiplying the w_i with $c(n)$ (see in eq. 5.24)

$$\widehat{MFCC}_l = w_i MFCC_i \tag{5.24}$$

5.7 Min Max Normalization

It is a technique that adjusts the coefficient in the range of [a, b]. Before the classification method, min-max normalization is applied to the features of both speech and ECG signals. Features are scaled into the range of [0, 1], and the following equation shows how to do it

$$X' = a + \frac{(X - X_{min})(b - a)}{X_{max} - X_{min}} \tag{5.25}$$

, wherein the proposed method b is equal to 1, and a is equal to 0.

5.8 Convolutional Neural Network

Before we begin with the CNN algorithm, the concept of a neural network, perceptron, multi-layer perceptron must be known. First, we begin with the idea of the neural network.

5.8.1 Neural Network

Neural network is a circuit of neurons or neuron simulations that form the structure and architecture of a human nervous system. It is, in a modern sense, called an artificial neural network composed of artificial neurons or nodes for solving artificial intelligence (AI) problems [59]. The neurons are connected by links, and they interact with each other. The nodes can take input data and perform a simple operation on that data. The result of these operations is then passed to other neurons. The output of each node is called **activation** or **node value**. Each link is associated with **weights**. In a nutshell, a single layer of the neural network is called **perceptron** and consists of 4 parts which are input values or one input layer, weights and bias, net sum, and activation function.

5.8.2 Introduction to Perceptron (Single Layer Network)

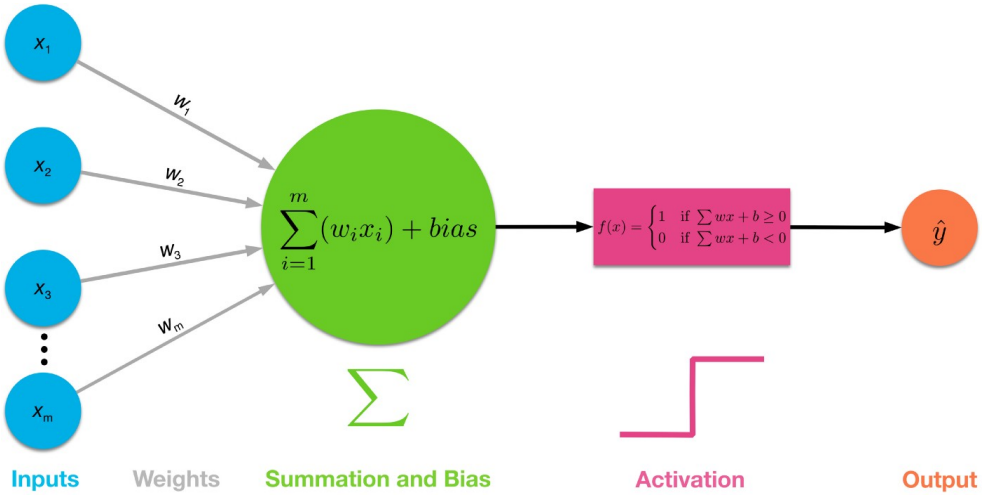


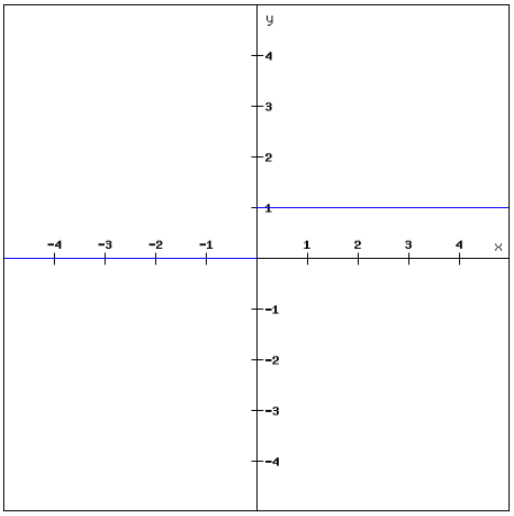
Figure 5.16: Procedures of single layer perceptron

Perceptron is a binary classification algorithm that was introduced by Cornell scientist Frank Rosenblatt. It helps for dividing a set of input signals into two parts which are "yes" or "no". It is a simple learning algorithm (uses binary step activation function) that weights the set of input how significant they are and generates an output decision of "0" and "1". Later, it has been developed further by the changes of the activation function. By the usage of linear activation function, the perceptron has developed for producing linear regression where it models the scalar response of the given input data. By the usage of the sigmoid activation function, the perceptron has developed for predicting the data with previously modeled data. It produces an output in the range of 0 to 1. the output will be close to 1 if the given data is close to previously modeled data. These are examples to understand the importance of activation function. Now, we will continue with the perceptron algorithm. A simple perceptron algorithm is given as

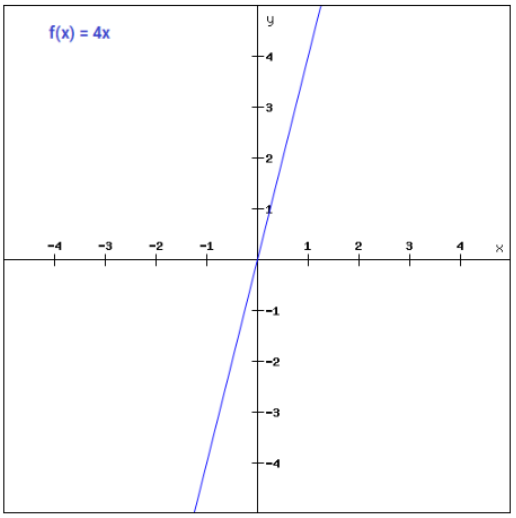
Transfer function of perceptron:

$$y = f\left(\sum_{i=1}^{m} w_i x_i\right) = f\left(w^T x\right) \tag{5.26}$$

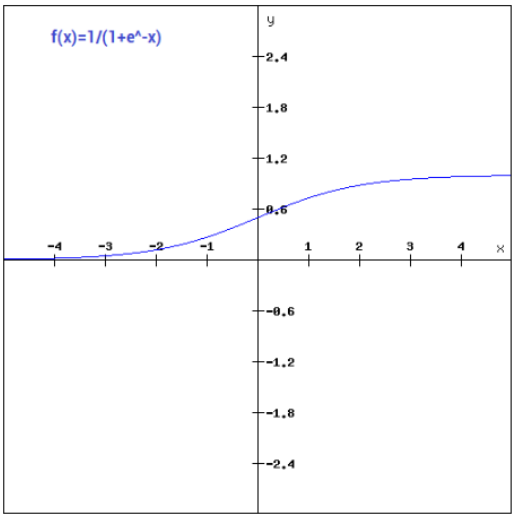
, where $[w_0, w_1, \cdots, w_m]$ and $x = [1, x_1, \cdots, w_m]$, and $f(x)$ is the activation function. In the equation, bias is represented with w_0 and for this reason x_0 always defined as 1 The activation function produces an output by combining the input to the neuron with weights. There are multiple activation functions such as binary step, linear function, etc.; however, the most used ones are sigmoid, and hyperbolic tangent, RELUs. The activation functions are given as



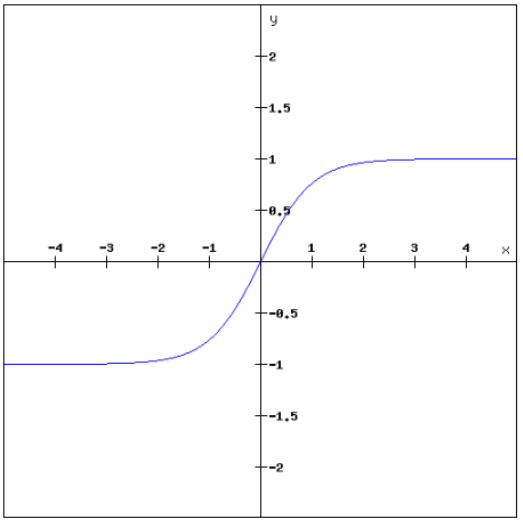
(a) Binary step function $f(x) = \begin{matrix} 1, x > 0 \\ 0, x \leq 0 \end{matrix}$



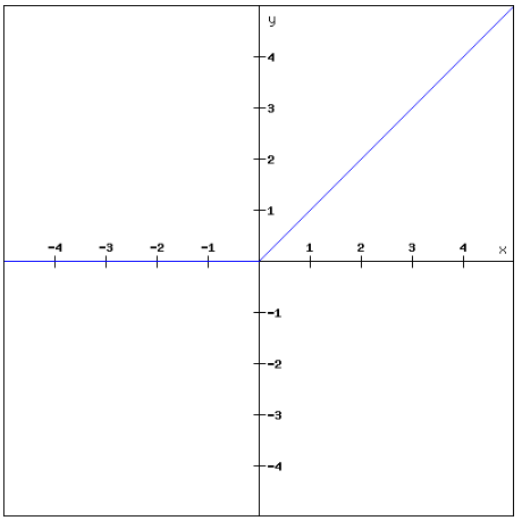
(b) Linear function $f(x) = ax$



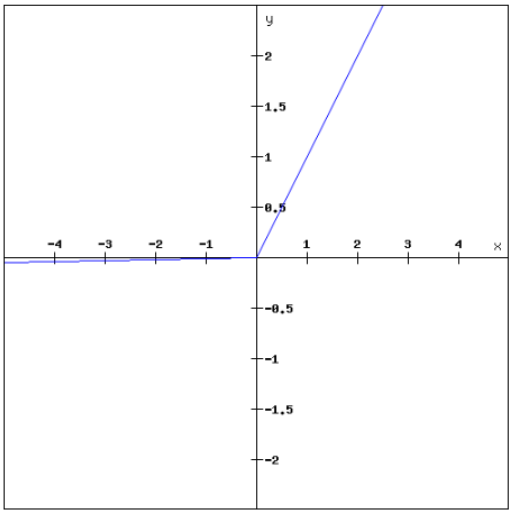
(c) Sigmoid function $f(x) = \frac{1}{1+e^{-x}}$



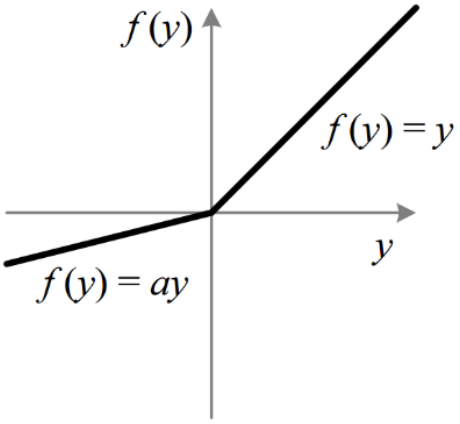
(d) Hyperbolic tangent function $f(x) = \frac{2}{1+e^{-2x}} - 1$



(e) RELU function $f(x) = \begin{matrix} x, x > 0 \\ 0, x \leq 0 \end{matrix}$



(f) Leaky RELU function $f(x) = \begin{matrix} x, x > 0 \\ 0.01x, x \leq 0 \end{matrix}$



(g) Parametrized RELU $f(x) = \begin{matrix} x, x > 0 \\ ax, x \leq 0 \end{matrix}$

Figure 5.17: Activation Functions

Let's define the sum S as

$$y = f\left(S\left(w_i, x_i\right)\right) \tag{5.27}$$

Mean Squared Error (Cost function): Learning occurs when changing (updating) the connection weights after each of data is processed, based on the amount of error in the output compared with expected result. The error (cost function) is given as

$$\varepsilon\left(w\right) = \frac{1}{2}\left(y - t\right)^2 \tag{5.28}$$

, where ε is the error, t is the target (class) values (from the training dataset), y is the algorithm's prediction for the training example using the activation function.

Gradient Descent Algorithm: This algorithm is used to train (or update) the weights of the perceptron to learn the given training dataset. This algorithm starts at an arbitrary position and iteratively converges to a value when error becomes minimum.

Using the gradient descent, the change in each weighs are defined as

$$w_i{}' = w_i - \mu \frac{\partial \varepsilon}{\partial w_i} \tag{5.29}$$

,where $w_i{}'$ is the weight after update, w_i is the weight before update, μ is learning rate (which is small constant)

Let's derive the error

$$\frac{\partial \varepsilon}{\partial w_i} = \frac{1}{2} \frac{\partial}{\partial w_i} (y - t)^2 \tag{5.30}$$

By using chain rule

$$(f \circ g)' = (f' \circ g) .g' \tag{5.31}$$

The equation 5.30 can be written as

$$\frac{\partial \varepsilon}{\partial w_i} = \frac{2}{2} (y - t) \frac{\partial}{\partial w_i} (y - t) = (y - t) \frac{\partial y}{\partial w_i} \tag{5.32}$$

,(where y is the only dependent variable of w)

Let's now calculate the derivative of y

$$\frac{\partial y}{\partial w_i} = \frac{\partial f (S (w_i, x_i))}{\partial w_i} \tag{5.33}$$

Once again, we use the chain rule to rewrite the equation 5.33

$$\frac{\partial f(S)}{\partial w_i} = \frac{\partial f(S)}{\partial S}\frac{\partial S}{\partial w_i} = x_i \frac{\partial f(S)}{\partial S} \tag{5.34}$$

The derivative of error becomes

$$\frac{\partial \varepsilon}{\partial w_i} = x_i (y - t) \frac{\partial f(S)}{\partial S} \tag{5.35}$$

Update Rule (General):

By merging 5.29 and 5.35, the weights can be updated with the following equation

$$w_i{}' = w_i - \alpha \frac{\partial \varepsilon}{\partial w_i} = w_i - \mu x_i (y - t) \frac{\partial f(S)}{\partial S} \tag{5.36}$$

In conclusion

$$w_i{}' = w_i - \mu x_i (y - t) \frac{\partial f(S)}{\partial S} \tag{5.37}$$

Update Rule (Linear Activation Function): By using the formula 5.37, update rule for linear activation function becomes

$$\varepsilon(w) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - t_i\right)^2 \tag{5.38}$$

$$w_k{}' = w_k + \mu [(y - t)] x_k, \ t \in [-\infty, \cdots, \infty] \tag{5.39}$$

$$w_0{}' = w_0 + \mu (y - t), \ t \in [-\infty, \cdots, \infty] \tag{5.40}$$

,where

$$y^{(i)} = w_i x_i \tag{5.41}$$

and t is the target value (which can be any real value that represents the input signals for linear function), k is the index of weight, m is the number of weight.

In conclusion, update rule for linear function becomes

$$w_k{}' = w_k - \mu \left(\sum_{i=1}^{m} w_i x_i - t_i \right) x_k,\ k \geq 1,\ t \in [-\infty, \cdots, \infty] \tag{5.42}$$

$$w_0{}' = w_0 - \mu \left(\sum_{i=1}^{m} w_i x_i - t_i \right),\ t \in [-\infty, \cdots, \infty] \tag{5.43}$$

5.38, 5.42 and 5.43 steps are repeated until ε converges.

Update Rule (Sigmoid Activation Function): The calculation of the sigmoid function is different from the linear activation function.

I will introduce 2 different assumptions for the sigmoid activation function. The first assumption is made by using maximum likelihood estimation.

Derivative of sigmoid function is

$$f'(z) = \frac{\partial}{\partial z} \left(\frac{1}{1+e^{-z}} \right) = \frac{1}{(1+e^{-z})^2} e^{-z} \tag{5.44}$$

$$= \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}} \right) \tag{5.45}$$

$$f'(z) = f(z)\,(1 - f(z)) \tag{5.46}$$

$$p(t = 1 \mid x; w) = f_w(x) \tag{5.47}$$

$$p(t = 0 \mid x; w) = 1 - f_w(x) \tag{5.48}$$

The more compact form can be written as

$$p(t \mid x; w) = (f_w(x))^t (1 - f_w(x))^{(1-t)},\ t\ \in \{0, 1\} \tag{5.49}$$

Likelihood relation is given as

$$L(w) = p(t \mid X; w) = \prod_{i=1}^{m} p\left(t^{(i)} \mid x^{(i)}, w\right) \tag{5.50}$$

$$\prod_{i=1}^{m} \left(f_w\left(x^{(i)}\right)\right)^{t^{(i)}} \left(1 - f_w\left(x^{(i)}\right)^{(1-t^{(i)})}\right) \tag{5.51}$$

By using equation 5.51, Log likelihood relation can be given

$$l(w) = \log L(w) = \prod_{i=1}^{m} t^{(i)} \log f_w\left(x^{(i)}\right) + \left(1 - t^{(i)}\right)\left[1 - \log f_w\left(x^{(i)}\right)\right] \tag{5.52}$$

Maximum log-likelihood by using gradient ascent is given as

$$w_i{}' = w_i + \mu \nabla_w l(w) \tag{5.53}$$

Minimize the $-l(w)$, (or maximize $l(w)$ is the same)

$$-l(w) = \sum_{i=1}^{m} \left[-t^{(i)} \log f_w\left(x^{(i)}\right) - \left(1 - t^{(i)}\right)\left[1 - \log f_w\left(x^{(i)}\right)\right]\right] \tag{5.54}$$

Write in different form (look eq. 5.26):

$$-l(w) = -t \log f_w\left(w^T x\right) - (1 - t)\left[1 - \log f_w\left(w^T x\right)\right] \tag{5.55}$$

Derivative of sum of terms:

$$-\frac{\partial}{\partial w_i} l(w) = -\frac{\partial}{\partial w_i} t \log f_w\left(w^T x\right) - \frac{\partial}{\partial w_i}(1 - t)\left[1 - \log f_w\left(w^T x\right)\right] \tag{5.56}$$

Derivative of $\log f_w(x)$:

$$= \left[-t \frac{1}{f_w\left(w^T x\right)} + (1 - t) \frac{1}{1 - f_w\left(w^T x\right)}\right] \frac{\partial}{\partial w_i} f_w\left(w^T x\right) \tag{5.57}$$

Chain Rule + derivative of f_w:

$$= \left[-t \frac{1}{f_w\left(w^T x\right)} + (1 - t) \frac{1}{1 - f_w\left(w^T x\right)}\right] f_w\left(w^T x\right)\left[1 - f_w\left(w^T x\right)\right] x_i \tag{5.58}$$

Algebraic Manipulation:

$$= \left[\frac{f_w\left(w^T x\right) - t}{f_w\left(w^T x\right)\left[1 - f_w\left(w^T x\right)\right]} \right] f_w\left(w^T x\right)\left[1 - f_w\left(w^T x\right)\right] x_i \tag{5.59}$$

$$= \left[f_w\left(w^T x\right) - t\right] x_i \tag{5.60}$$

It is cosmetically identical to the Mean Squared Error Rule. However, $f_w(x)$ is non-linear in here.

Update Rule becomes (cross entropy cost function):

$$C(w) = -\frac{1}{2m} \sum_{i=1}^{m} \left[t_i \log f_w(w_i x_i) + (1 - t_i) \log\left(1 - f_w(w_i x_i)\right) \right], \; t \in \{0, 1\} \tag{5.61}$$

$$w_i{}' = w_i - \mu \frac{\partial l(w)}{\partial w_i} \tag{5.62}$$

$$w_k{}' = w_k - \mu \left(\frac{1}{1 + e^{-\left(\sum_{i=1}^{m} w_i x_i\right)}} - t_i \right) x_k, \; k \geq 1, \; t \in \{0, 1\} \tag{5.63}$$

$$w_0{}' = w_0 - \mu \left(\frac{1}{1 + e^{-\left(\sum_{i=1}^{m} w_i x_i\right)}} - t_i \right), \; t \in \{0, 1\} \tag{5.64}$$

,where

$$f_w(w_i x_i) = \frac{1}{1 + e^{-(w_i x_i)}} \tag{5.65}$$

and t is the class number (which can be 0 or 1 for sigmoid function), k is the index of weight, m is the number of weight. 5.61, 5.63, 5.64 and 5.65 steps are repeated until ε converges.

Second assumption is (by using mean squared error):

$$\varepsilon(w) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - t_i \right)^2, \; t \in \{0, 1\} \tag{5.66}$$

$$\frac{\partial \varepsilon(w)}{\partial w_k} = \sum_{i=1}^{m} \left(y^{(i)} - t_i \right) f'(w_i x_i) \, x_k \tag{5.67}$$

By using the 5.46, it becomes

$$\frac{\partial \varepsilon(w)}{\partial w_k} = \sum_{i=1}^{m} \left(y^{(i)} - t_i\right) f(w_i x_i) (1 - f(w_i x_i)) x_k \tag{5.68}$$

Update rule:

$$w_k{}' = w_k \mu \frac{\partial \varepsilon(w)}{\partial w_k} \tag{5.69}$$

$$w_k{}' = w_k - \mu \sum_{i=1}^{m} \left(y^{(i)} - t_i\right) f(w_i x_i) (1 - f(w_i x_i)) x_k, \ k \geq 1, \ t \in \{0, 1\} \tag{5.70}$$

$$w_0{}' = w_0 - \mu \sum_{i=1}^{m} \left(y^{(i)} - t_i\right) f(w_i x_i) (1 - f(w_i x_i)), \ t \in \{0, 1\} \tag{5.71}$$

Update Rule (RELU function):

RELU updating rule is the same as linear function when $x > 0$; however, whenever the value of x has become 0 or smaller than zero, weights updating stops because of derivative of RELU is undefined for $x \leq 0$. Because the gradient becomes zero whenever x becomes lower than zero, the weight stop updates. RELU can be fragile during the large number of training data set and can irreversibly "die" (which can't update weights). If the learning rate is set too high, we may cross the large number of "dead" neurons that can never be activated again. However, the speed of RELU is much higher than other activation functions, and it reduces the likelihood of the gradient vanishing. **Vanishing gradients** lead to very small changes in the weights proportional to the partial derivative of the error function. As more layers using certain activation functions (e.g., sigmoid, tanh) are added to the neural network, the gradient of the loss function approaches zero, making the network hard to train. To solve the problem, RELU activation functions are used.

Update Rule (Leaky RELU): To solve dead neurons, leaky RELU is proposed.

$$w_k{}' = w_k - \mu \sum_{i=1}^{m} (w_i x_i - t_i) x_k, \ k \geq 1 \ and \ x_k > 0 \tag{5.72}$$

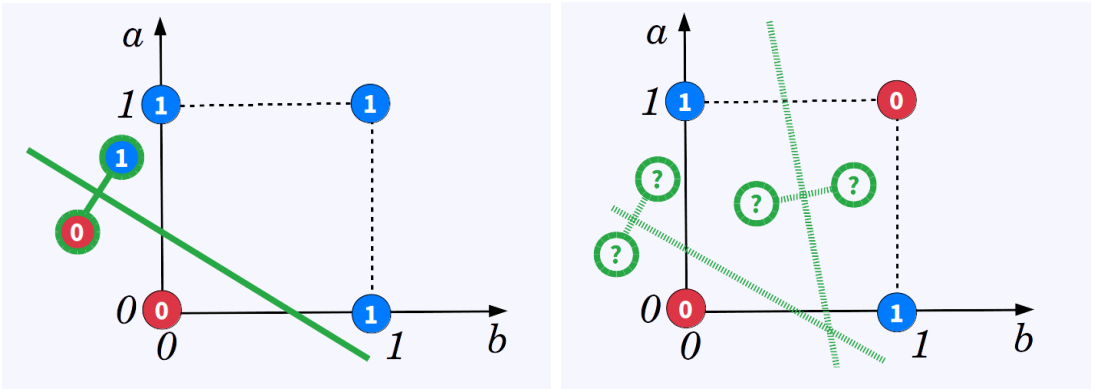
$$w_k{}' = w_k - 0.01 \left[\mu \sum_{i=1}^{m} (w_i x_i - t_i)\, x_k \right],\ k \geq 1\ and\ x_k \leq 0 \tag{5.73}$$

$$w_0{}' = w_0 - \mu \sum_{i=1}^{m} (w_i x_i - t_i) \tag{5.74}$$

If the value of 0.01 changes to α, it will become parametrized RELU.

5.38, 5.41, 5.42, 5.43 and 5.44 steps are repeated until ε converges.

By using the sigmoid activation function, perceptron becomes logistic regression. The logistic regression is used because of tackling multi-classification problems, e.g., the using One-vs-All or One-vs-One approaches via the related Softmax regression or multinomial logistic regression. Although there are kernelized variants of logistic regression exist, the standard "model" is a linear classifier. Thus, logistic regression is useful when the classes in the dataset are more or less "linearly" separable. However, the classic perceptron is not enough for handling the supervised classification problem. We know that perceptron is a linear classifier algorithm that classifies the input by separating two categories with a line. The problem of classic perceptron algorithm is that input of classes can't exhibit two different traits such as XOR function (XOR operator trigger when input exhibits either one trait or another, but not both; it stands for "exclusive OR"). Classic perceptron can't perform classification on non-linear inputs. The problem is shown in the following figures (see figure 5.18(a), figure 5.18(b))



(a) The space of the OR function

(b) The space of the XOR function

Figure 5.18: Response of perceptron to non-linear inputs

a	b	y=a + b
0	0	0
0	1	1
1	0	1
1	1	1

The space of the OR function can be drawn where X and Y axis are respectively "a" and "b" inputs. The green line is the separation line where $y = 0$. Perceptron can find an optimal solution when transfer function of perceptron is given as (linear activation)

$$y = w_0 + w_1 a + w_2 b$$

a	b	y=a $\bigoplus$ b
0	0	0
0	1	1
1	0	1
1	1	0

The space of the XOR function is illustrated. Unfortunately, the perceptron is not able to discriminate zeros from ones.

To solve the single perceptron problem, a multi-layer perceptron is introduced, which capable of approximating an XOR operator as well as many other non-linear functions.

5.8.3 Multilayer Perceptron (Neural Network)

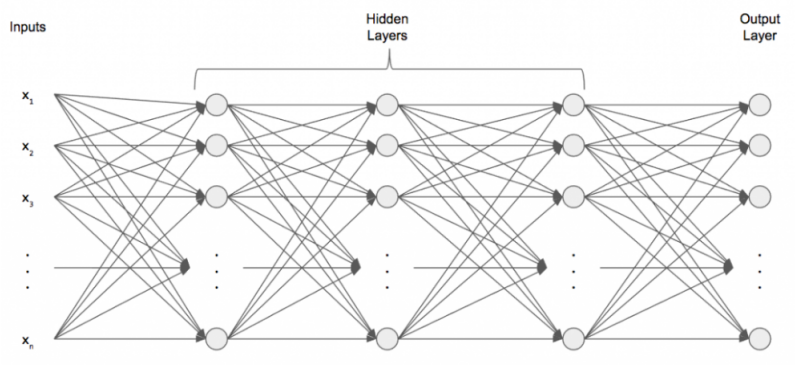


Figure 5.19: Multilayer Perceptron

The multi-layer perceptron (MLP) is composed of more than one perceptron where each one of perceptron has an input layer to receive signal, an output layer to make the decision or prediction about input layer, and the arbitrary number of hidden layers between inputs and output layers which are the true computational engine of the MLP. Architecture of multilayer perceptron is given in figure 5.19. It can be seen that each output of the perceptron is weighed of the other perceptron. This continues until the final output layer is constructed. The calculation now differs with a single perceptron. There will be two concepts which are feed-forward and backpropagation. However, before we start with the concept, we need to understand the notation of multi-layer perceptron. (Notation will be different with respect to single perceptron; however, it does the same job.)

$$z = \sum_{i=1}^{N} W_i x_i = W^T x \tag{5.75}$$

$$\alpha = f(z) \tag{5.76}$$

,where z will called **pre-activation** and α will called **post-activation**. This equation is similar to single perceptron. The difference is now:

$$z_k = \sum_{j} W_{kj} x_k \tag{5.77}$$

$$a_k = f(z_k) \tag{5.78}$$

The input layer is indexed as j and the output layer by k. The formula is described of getting the pre-activation of an arbitrary neuron k in the output layer. Each input multiply by weights that connect that input to k^{th} neuron. (see in figure 5.20). That means the first output neuron is $k = 1$. Component form of the figure 5.20 will be useful when we discuss backpropagation (weight updates).

α is vector which represents output of each activation (post-activations), W is weight matrix, x is the input vector

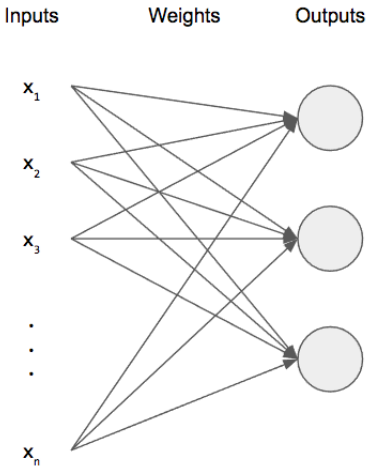


Figure 5.20: Architecture of One-layer Multilayer Perceptron

Now, we need a notation for representing each hidden layer (see in figure 5.19).

$$z^{(l)} = W^{(l)}\alpha^{(l-1)} \tag{5.79}$$

$$\alpha^{(l)} = f\left(z^{(l)}\right) \tag{5.80}$$

,where $\alpha^{(l)} = x$ and $l \in [2, L]$, L is the maximum number of hidden layer. Now, you should notice that in the pre-activation, we no longer just referring to x, rather we take the activation of previous layer $l - 1$ when computing the pre-activation of layer l

Now cost function will be

$$C(W) = \frac{1}{2} \sum_{x \in X} \left\| y(x) + \alpha^{(L)}(x) \right\|^2 \tag{5.81}$$

$$W_{jk}^{(l)'} = W_{jk}^{(l)} - \mu \frac{\partial C}{\partial W_{jk}^{(l)}} \tag{5.82}$$

,where W represents all the weights of network, the sum is over all training example, $\alpha^{(L)}$is the output layer vector when given $f(x)$ vector.

Feedforward Process for multi-layer perceptron:

$$\alpha_j^{(l)} = f\left(\sum_k W_{jk}^{(l)} \alpha_k^{(l-1)} \right) \tag{5.83}$$

Now think that we have 100.000 vectors in a dataset which represent input signals. For speeding the algorithm, the randomly 1000 training data are selected. However, we need to make sure that all the data in the dataset are given for training. For this reason, data are divided into mini-batches (which are 100 vector sets). At each mini-batch, the average gradient and parameters are updated by the move on the next mini-batch. After all the mini-batch is finished, all the data is shuffled and again divided into mini-batch. This process is called one **epoch**. Now, the algorithm is needed to stop at some point. For this reason, one of the mini-batches is used to calculate the performance of the system. If the performance is enough for describing the dataset, then parameter updating will be stopped.

Let's continue with the feed-forward process:

In the feed-forward process, initial weights are chosen as small randomly constant and find the outputs of all neurons in the first layer by using initial weights. Then output produced by the first layer is given to the second layer as input, and output of the second layer is calculated. Then continue until the final output is found, which represents the output of multi-layer perceptron. The equation is

represented in (5.83). From now on, the weight parameters of each neuron are calculated by taking the partial derivative of the cost function produced by each neuron.

Backpropagation Process for multi-layer perceptron:

Now, we need to introduce new notation which is error $\delta_j^{(l)}$ of neuron j in layer l by

$$\delta_j^{(l)} = \frac{\partial C}{\partial z_j^{(l)}} \tag{5.84}$$

, where $\delta_j^{(l)}$ is the error vector associated with layer l

An equation for the error in the output layer is given as:

$$\delta_j^{(L)} = \frac{\partial C}{\partial \alpha_j^{(L)}} f'\left(z_j^{(L)}\right) \tag{5.85}$$

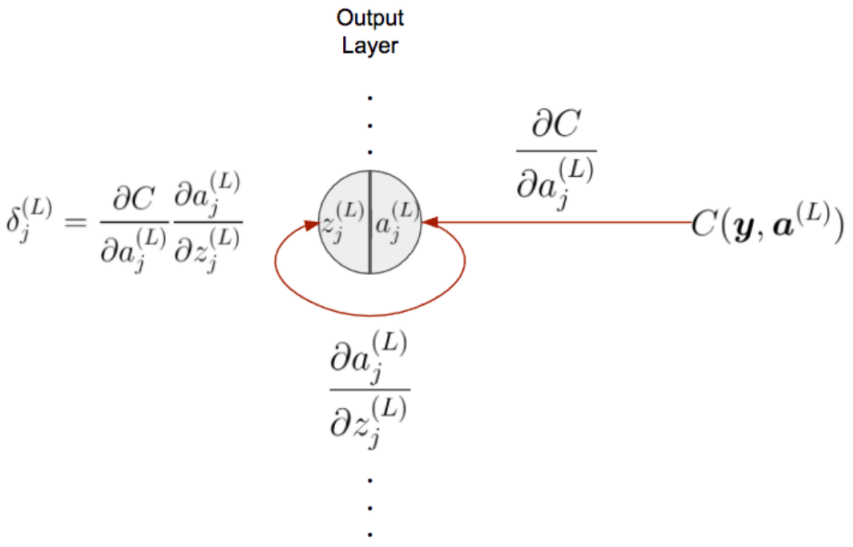


Figure 5.21: Backpropagation (input-output relation of final neuron)

Eq. 5.85 is the first equation of backpropagation. It represents the local error gradient at the last layer. Now, we have the last layer's error gradient, we need an equation that tells us how to propagate that error gradient backward.

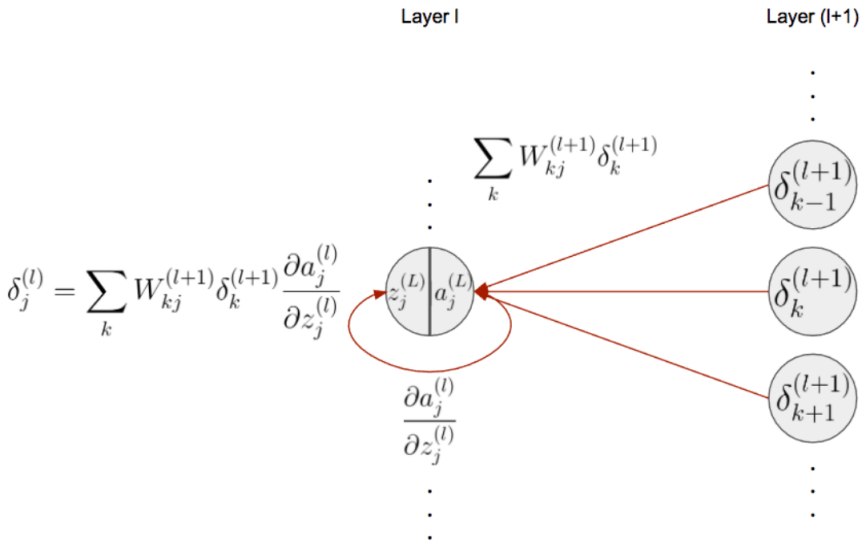


Figure 5.22: Backpropagation (input-output relation of hidden neuron)

Figure 5.22 is the extension of backpropagation which adding layers. The equation becomes considering all the hidden layer.

$$\delta_j^{(L)} = \sum_k W_{jk}^{(l+1)} \delta_k^{(l+1)} f'\left(z_j^{(l)}\right) \tag{5.86}$$

Now, we can use the local error gradients to calculate partial derivatives of the weights at a given layer. The equation is given as

$$W_{jk}^{(l)\prime} = W_{jk}^{(l)} - \mu \frac{1}{m} \sum_{x_{(1)},\cdots,x_{(m)}} \frac{\partial C_x}{\partial W_{jk}^{(l)}} \tag{5.87}$$

$$\frac{\partial C}{\partial W_{jk}^{(l)}} = \delta_i^{(l)} \alpha_k^{(l-1)}, \ k \geq 1 \tag{5.88}$$

$$\frac{\partial C}{\partial W_{j0}^{(l)}} = \delta_i^{(l)} \tag{5.89}$$

In summary, the following algorithm is shows the gradient descent learning step based on mini batch:

1. Input a set of training example

2. For each training example x, set the corresponding input activation $\alpha^{(x,1)}$ and perform the following steps:

   - Feedforward: for each layer $l = 2, 3, \cdots, L$ compute the vector $z^{(x,l)} = w^{(l)}\alpha^{(x,l-1)}$ and $\alpha^{(x,l)} = f\left(z^{(x,l)}\right)$
   - Output error : compute the following equation

$$\delta_j^{(x,L)} = \frac{\partial C}{\partial \alpha_j^{(x,L)}} f'\left(z_j^{(x,L)}\right)$$

   - Back propagate the error: for each $l = L, L-1, \cdots, 2$ update the weights according the following rule

$$w^{(l)\prime} = w^{(l)} - \mu \frac{1}{m} \sum_{x_{(1)},\cdots,x_{(m)}} \delta_j^{(x,l)} \left(\alpha^{x(l-1)}\right)^T$$
$$w_0^{(l)\prime} = w_0^{(l)} - \mu \frac{1}{m} \sum_{x_{(1)},\cdots,x_{(m)}} \delta_j^{(x,l)}$$

To further your understanding of multi-layer perceptron and neural network, please look at the Neural Networks and Deep Learning book of Michael Nielsen [60].

5.8.4 Softmax Activation Function for Classification

This function is used to solve multi class problem where we have multiple class names. Softmax activation function is applied in the end of hidden layer so that it can separate the classes. The function can be given as

$$a_j^{(L)} = \frac{e^{x_j^{(L)}}}{\sum_k e^{z_k^{(L)}}} \tag{5.90}$$

, where $a_j^{(L)}$ is the activation of j^{th} output neuron

Lets' look at the Softmax algorithm:

Softmax function is used when the target value t can take the values from the discrete set $\{1, 2, \cdots, k\}$. In that case, we assume

$$\begin{aligned} p(t=1; x; w) &= \phi_1 \\ p(t=2; x; w) &= \phi_2 \\ &\vdots \\ p(t=k-1; x; w) &= \phi_{k-1} \\ p(t=k; x; w) &= 1 - \textstyle\sum_{i=1}^{k-1} \phi_i = \phi_k \end{aligned} \tag{5.91}$$

In order to estimate the parameters of classifier of k classes, $k-1$ parameter must be estimated (the last one determined by the other ones like binary classifier). Now it can be written as

$$p(t; x; w) = p(t=1; x; w)^{I\{t=1\}} p(t=2; x; w)^{I\{t=2\}} \cdots p(t=k; x; w)^{I\{t=k\}} \tag{5.92}$$

, where $I\{\}$ indicates a function if it is true ,then return 1, otherwise return 0

Now, we will estimate ϕ's with the following equation

$$\begin{aligned} \widehat{\phi}_1 &= f_{w^{(1)}}(x) = \frac{exp\left(w^{(1)}\right)^T x}{\sum_{i=1}^{k} exp\left(w^{(i)}\right)^T x} \\ \widehat{\phi}_2 &= f_{w^{(2)}}(x) = \frac{exp\left(w^{(2)}\right)^T x}{\sum_{i=1}^{k} exp\left(w^{(i)}\right)^T x} \\ &\vdots \\ \widehat{\phi}_{k-1} &= f_{w^{(k-1)}}(x) = \frac{exp\left(w^{(k-1)}\right)^T x}{\sum_{i=1}^{k} exp\left(w^{(i)}\right)^T x} \\ \widehat{\phi}_k &= f_{w^{(k)}}(x) = \frac{exp\left(w^{(k)}\right)^T x}{\sum_{i=1}^{k} exp\left(w^{(i)}\right)^T x} \end{aligned} \tag{5.93}$$

,where $w^{(i)}$ for all $i = 1, \cdots, k-1$, and $w^{(k)} = \overrightarrow{0}$ are the parameter that are used to generate the hypothesis. Let θ be the parameter matrix that contains $w^{(i)}$ for

all $i = 1, \cdots, k-1$

$$\theta = \begin{bmatrix} \vdots & \vdots & & \vdots & \vdots \\ w^{(1)} & w^{(2)} & \cdots & w^{(k-1)} & w^{(k)} \\ \vdots & \vdots & & \vdots & \vdots \end{bmatrix} \tag{5.94}$$

Note that $\sum_{i=1}^{k} \widehat{\phi}_i = 1,\ 0 \leq \widehat{\phi}_i \leq 1$ so the $w^{(1)}$'s form a probability distribution.

Now, let's see maximum likelihood criteria on log likelihood function. It is given as

$$l(\theta) = \log p\left(\overrightarrow{t} \mid X; \theta\right) \tag{5.95}$$

$$= \log p\left(t^{(1)}, \cdots t^{(n)} \mid x^{(1)}, \cdots, x^{(n)}; \theta\right) \tag{5.96}$$

$$= \log \prod_{i=1}^{n} p\left(t^{(i)} \mid x^{(i)}; \theta\right) \tag{5.97}$$

$$= \log \prod_{i=1}^{n} f_{w^{(1)}}\left(x^{(i)}\right)^{I\left\{t^{(i)}=1\right\}} f_{w^{(2)}}\left(x^{(i)}\right)^{I\left\{t^{(i)}=2\right\}} \cdots f_{w^{(k)}}\left(x^{(i)}\right)^{I\left\{t^{(i)}=k\right\}} \tag{5.98}$$

$$= \log \sum_{i=1}^{n} I\left\{t^{(i)}=1\right\} \log f_{w^{(1)}}\left(x^{(i)}\right) + \cdots + I\left\{t^{(i)}=k\right\} \log f_{w^{(k)}}\left(x^{(i)}\right) \tag{5.99}$$

$$= \log \sum_{i=1}^{n} \sum_{q=1}^{k} I\left\{t^{(i)}=q\right\} \log f_{w^{(q)}}\left(x^{(i)}\right) \tag{5.100}$$

,where in 5.97, we used the assumption which the training set was generated independently, in equation 5.98, 5.99, we used the logarithm properties, and in equation 5.100, we simplified the expression. Let's substitute the hypothesis with its explicit function to get

$$= \log \sum_{i=1}^{n} \sum_{q=1}^{k} I\left\{t^{(i)}=q\right\} \log f_{w^{(q)}}\left(x^{(i)}\right) \tag{5.101}$$

$$= \sum_{i=1}^{n} \sum_{q=1}^{k} I\left\{t^{(i)}=q\right\} \log \frac{exp(w^{(q)})^T x^{(i)}}{\sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)}} \tag{5.102}$$

$$= \sum_{i=1}^{n} \sum_{q=1}^{k} I\{t^{(i)} = q\} \left[\log exp(w^{(q)})^T x^{(i)} - \log \sum_{p=1}^{k} exp(w^{(p)})^T x^{(i)} \right] \tag{5.103}$$

$$= \sum_{i=1}^{n} \left[\sum_{q=1}^{k} I\{t^{(i)} = q\} \left(w^{(q)}\right)^T x^{(i)} - \sum_{q=1}^{k} I\{t^{(i)} = q\} \log \sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)} \right] \tag{5.104}$$

$$= \sum_{i=1}^{n} \left[\sum_{q=1}^{k} I\{t^{(i)} = q\} \left(w^{(q)}\right)^T x^{(i)} - \log \sum_{p=1}^{k} exp(w^{(p)})^T x^{(i)} \right] \tag{5.105}$$

,where in equation 5.102, 5.103, logarithm properties were used and in equation 5.104, we used the fact that the indicator function return 1 only once for every training example. Now, lets' continue to find derivatives of $l(\theta)$ with respect to $w^{(j)}$ to form the gradient ascent update rule which will maximize the log likelihood function

$$\frac{\partial l(\theta)}{\partial w_j^{(r)}} = \frac{\partial}{\partial w_j^{(r)}} \sum_{i=1}^{n} \left[\sum_{q=1}^{k} I\{t^{(i)} = q\} \left(w^{(q)}\right)^T x^{(i)} - \log \sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)} \right] \tag{5.106}$$

$$= \sum_{i=1}^{n} \left[\sum_{q=1}^{k} \frac{\partial}{\partial w_j^{(r)}} I\{t^{(i)} = q\} \left(w^{(q)}\right)^T x^{(i)} - \frac{\partial}{\partial w_j^{(r)}} \log \sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)} \right] \tag{5.107}$$

$$= \sum_{i=1}^{n} \left[I\{t^{(i)} = r\} x_j^{(i)} - \frac{exp\left(w^{(r)}\right)^T x^{(i)}}{\sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)}} x_j^{(i)} \right] \tag{5.108}$$

$$= \sum_{i=1}^{n} \left[I\{t^{(i)} = r\} - \frac{exp\left(w^{(r)}\right)^T x^{(i)}}{\sum_{p=1}^{k} exp\left(w^{(p)}\right)^T x^{(i)}} \right] x_j^{(i)} \tag{5.109}$$

$$= \sum_{i=1}^{n} \left[I\{t^{(i)} = r\} - p\left(t^{(i)} = r \mid x^{(i)}; \theta\right) \right] x_j^{(i)} \tag{5.110}$$

, where in equation 5.107, we use linearity of the derivative, in equation 5.108, we took derivatives, in equation 5.109, 5.110 we simplified the expression. Finally, the batch gradient ascent update rule becomes

$$w_j^{(r)\prime} = w_j^{(r)} + \mu \sum_{i=1}^{n} \left[I\{t^{(i)} = r\} - p\left(t^{(i)} = r \mid x^{(i)}; \theta\right) \right] x_j^{(i)} \tag{5.111}$$

Note that the gradient's value goes to 0 when

$$t^{(i)} = r,\ p\left(t^{(i)} = r \mid x^{(i)}; \theta\right) \rightarrow 1 \tag{5.112}$$

$$t^{(i)} \neq r,\ p\left(t^{(i)} = r \mid x^{(i)}; \theta\right) \rightarrow 0,\ \forall_i,\ r \tag{5.113}$$

So the gradient stops adjusting θ when is close to 1 for the correct class and $t^{(i)} = r,\ p\left(t^{(i)} = r \mid x^{(i)}; \theta\right)$ close to 0 for the incorrect classes, that approves the that the results make sense.

For more information, please look at the machine learning lecture of Andrew NG [61].

5.8.5 Understanding of Convolutional Neural Network

Parameter updating of multi-layer perceptron consumes much time in the process. The cost of fitting the parameters is increasing when the training dataset is getting bigger. For this reason, a new type of system is needed to minimize the features in the training data. The system architecture is given in figure 5.23.

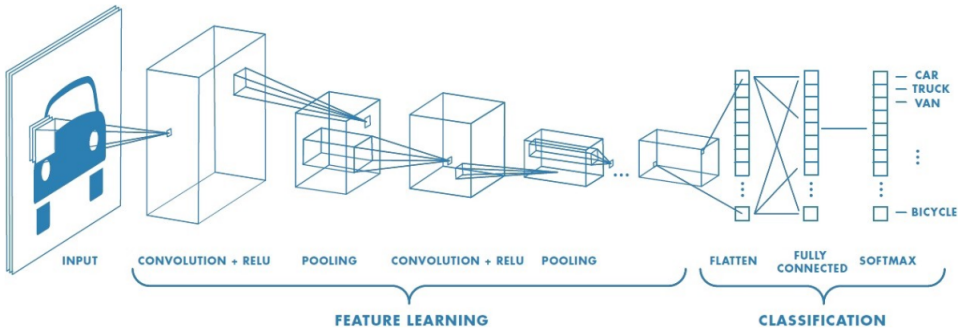


Figure 5.23: Convolutional Neural Network [62]

CNN has two parts which are feature learning and classification. In feature learning, there is no parameter updating; it finds the feature for the input by applying multiple convolutional layers and RELU operation after each convolutional layer to delete the negative part of the input. Pooling operations are used to minimize

the size of the input. These operations continue until the flatten layer. In the flatten layer, all 2-D images are mapped into 1-D vectors. Now, we are in the classification part. In this part, values in the flatten layers are given to multi-layer perceptron in batches. Now, there are two sections. In the first section, learning parameters are found by using the training dataset. In the second section, classification occurs in the given test dataset. The softmax activation function is used on the output neuron of the multi-layer perceptron to do the classification. This activation is applied because it finds the given input's probabilities for each class, and it puts the given input to a class where the probability is close to 1.

The classification algorithm is explained previously. In this section, we will dive into feature learning algorithms (layers) step by step.

5.8.5.1 Convolutional Layer

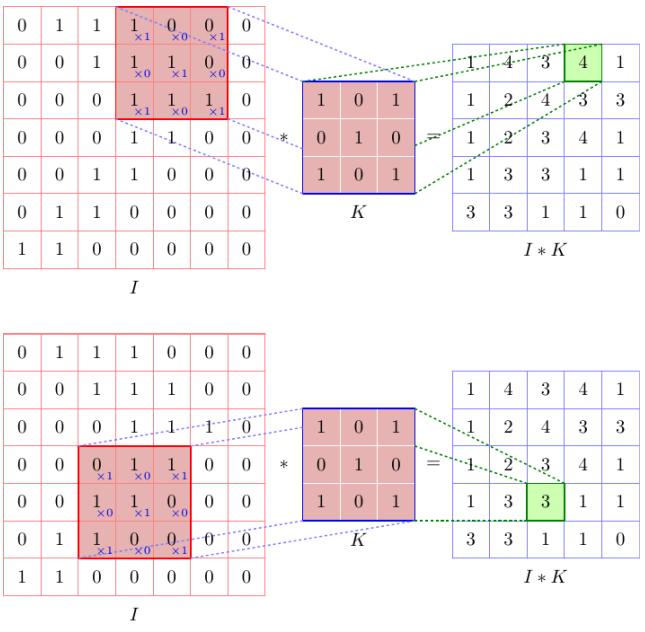


Figure 5.24: 2-D Convolution

It is a process where the input signal is passed through a filter (in other words, kernel). In most cases, images are used as input signals, and 2-D filters are used for the operations. 2-D convolutional output is calculated with the following

equation:

$$G[m,n] = (g * h) = \sum_j \sum_k h[j,k]\, g[m-j, n-k] \tag{5.114}$$

, where input image denoted by g and kernel by h

In this section, there are multiple options. If the output size of the convolution wants to be the same as the input size, then zero pads will need to add to the borders of the image. Padding width should meet the following equation, where p represents padding, h is the filter dimension (usually odd number)

$$p = \frac{h-1}{2} \tag{5.115}$$

For some cases, to reduce the resolution of the input image pixel, a stride operation is applied to the convolution.

$$G[m,n] = \sum_j \sum_k h[j,k]\, g[ms_1 - j, ns_2 - k],\ m \leq \frac{M}{s_1},\ n \leq \frac{N}{s_2} \tag{5.116}$$

,where M is the number of image pixel in the row, N is the number of image pixel in the column, s_1 is the stride for rows, s_2 is the stride for column. Assuming that $\frac{M}{s_1}$ is and $\frac{N}{s_2}$ gives us integer.

Now if we take into account of stride and padding operations, the dimension of the output matrix is calculated by the following equation.

$$n_{out} = \frac{n_{in} + 2p - h}{s} + 1 \tag{5.117}$$

,where $s_1 = s_2 = s,\ m = n$

The dimension calculation is important because the next step calculation is done by considering the input dimension. After every convolutional layer, RELU operation is applied pixel by pixel. This operation eliminates the negative values of the output image.

Note: 2-D convolutional layer are used when the data points in the matrix are important. If it is not, the signal can be mapped into 1-D vectors, and a 1-D filter operation can be done.

5.8.5.2 Pooling Operation

The pooling operation involves sliding a 2-D filter over each channel of the feature map and summarizing the features lying within the region covered by the filter. Pooling operation is used to reduce the dimensions of feature maps. We will discuss two types of pooling operation, which are max pooling and average pooling.

Max pooling: It is a pooling operation that selects the maximum element from the region of the feature map covered by the filter (see in figure 5.25).

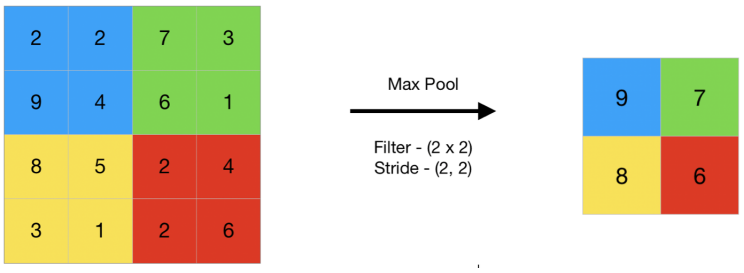


Figure 5.25: Max pooling Operation

Average pooling: It computes the average of the elements in the region of the feature map covered by the filter (see in figure 5.26).

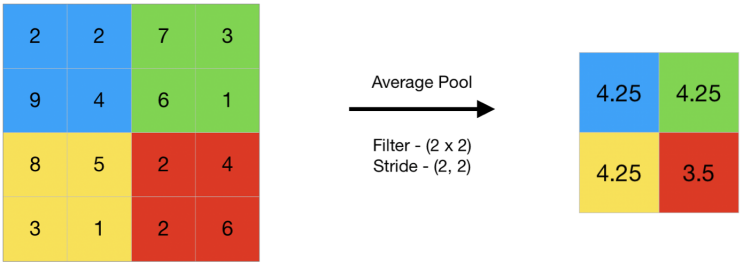


Figure 5.26: Average pooling Operation

Feature dimension (n_h x n_w x n_c) after the pooling operation can be given as

$$n_{out} = \frac{n_h - h + 1}{s}, \quad n_{out(w)} = \frac{n_w - h + 1}{s} \quad n_{out(c)} = n_c \tag{5.118}$$

, where n_h the height of the feature map , n_w is the width of the feature map, n_c is the number of channel in the feature map, h is filter size, s is the stride length

5.8.5.3 Flatten Layer

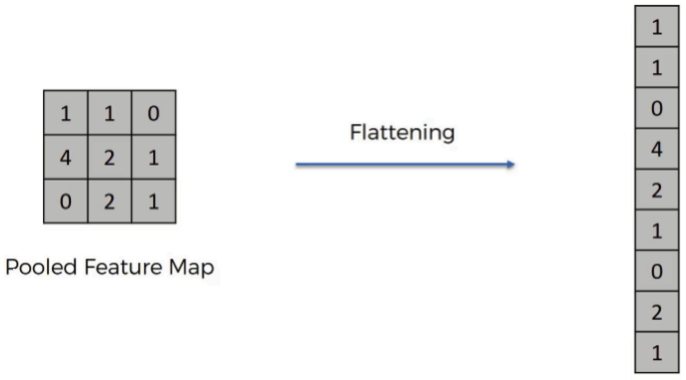


Figure 5.27: Average pooling Operation

This layer is the input layer for the classification part. It sorts the matrix into a 1-D vector which represents the input feature of the multi-layer perceptron.

5.8.5.4 Fully Connected Layer

It represents the hidden layer in the multi-layer perceptron. In the CNN, there can be multiple fully connected layers, and linear, RELU, leaky RELU, or sigmoid activation functions can be given for each layer. In most cases, the RELU activation function is used because of processing speed. The most important part is, number of output neurons for the final fully connected layer must be the same as the number of classes. In the final fully connected layer, the Softmax activation function needs to be used, which predicts the given inputs class numbers.

5.8.5.5 Understanding of Output Size of Each Layer

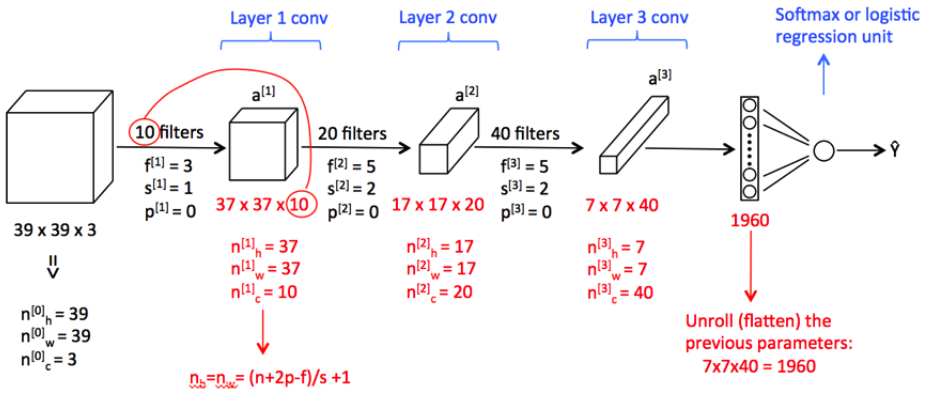


Figure 5.28: 3 layers CNN

In the figure 5.28, 3 layers CNN is illustrated. Input image size is given as (39 x 39 x 3) three-dimensional matrix in which the first two dimensions indicate rows and columns image pixels, whereas the final dimension indicates its RGB colors. In the first 2-D convolutional layer, 10 filters in the filter bank, which corresponds to different shapes, are applied to the image. Filter dimensions "f" are defined as (3 x 3), and stride "s" is defined as 1, and padding "p" is defined as 0 in the first layer. Zero padding means there is no padding applies to the image. By using the equation 5.117, the dimension of the feature map for the first layer can be found. "Note that each output feature map is the sum of convolutions of all the input feature map" [63]. That means all the channel "n_c" of the input image or feature map are summed together, then filters in the filter bank are applied onto them so that output channel size becomes the number of filter size. That is the reason why the output channel size of the feature map is the same as the number of filters. The dimension calculation is also the same with the second and third convolution layers. It differs in the flatten layer, which sorts the (7 x 7 x 40) matrix into a 1-D element. From that forward, classification occurs, which is explained previously.

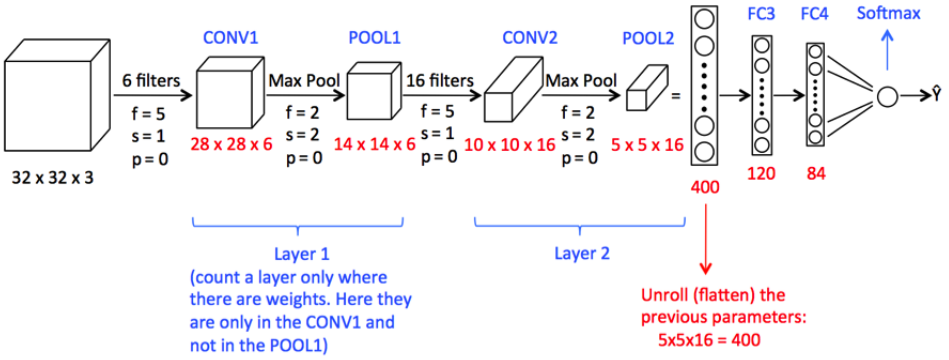


Figure 5.29: 2 layer CNN with max pooling operation

In figure 5.29, dimension calculation of feature map for each convolution layer are the same. It differs in pooling operation. This time, equation 5.118 is applied to find the filter size. The output size of the first pooling operation can be given as

$$n_{out(h)} = \frac{28-2+1}{2} = 13.5,\ n_{out(w)} = \frac{28-2+1}{2} = 13.5,\ n_{out(c)} = 6 \text{ filters} \tag{5.119}$$

Output size of the second pooling operation can be given as

$$n_{out(h)} = \frac{14-5+1}{1} = 10,\ n_{out(w)} = \frac{14-5+1}{1} = 10,\ n_{out(c)} = 16 \text{ filters} \tag{5.120}$$

Output feature map is optimized if the dimension of pooling operation does not give us integer. It converges the dimension of the output feature map to the nearest value, which is 14.

5.8.5.6 CNN Architecture of Proposed Algorithm

CNN's are the feed-forward ANN with additional convolutional and subsampling layers [64]. By selecting the appropriate convolutional layer, hidden layers, we can either train a massive 2-D visual database or 1-D speech or ECG database with a proper training process. In our research, we focused on the 1-D CNN approach, where has four convolutional layers with the same padding attitude

and 2 fully connected layers. The architecture has shown in the following table, where N represents the number of classes.

Table 5.1: 1-D CNN architecture

1-D CNN architecture for ECG Based identification		**1-D CNN architecture for Speech Based identification**	
Layer	**Size**	**Layer**	**Size**
Input Layer	(1x485)	Input Layer	(1x20)
Convolutional Layer	(1x3, 2)	Convolutional Layer	(1x9, 8)
batchNormalization Layer	-	batchNormalization Layer	-
Relu Layer		Relu Layer	
Convolutional Layer	(1x5, 4)	Convolutional Layer	(1x12, 16)
batchNormalization Layer	-	batchNormalization Layer	-
Relu Layer		Relu Layer	
Convolutional Layer	(1x7, 8)	Convolutional Layer	(1x15, 32)
batchNormalization Layer	-	batchNormalization Layer	-
Relu Layer		Relu Layer	
Convolutional Layer	(1x9, 16)	Maxpooling Layer	(1x2)
batchNormalization Layer	-	Convolutional Layer	(1x18, 64)
Relu Layer		batchNormalization Layer	-
Maxpooling Layer	(1x2)	Relu Layer	
Fully Connected Layer	1x(485*N)	Maxpooling Layer	(1x2)
Fully Connected Layer	1 x N	Fully Connected Layer	1x(20*N)
Softmax Layer	1 x N	Fully Connected Layer	1 x N
		Softmax Layer	1 x N

Chapter 6

Approach

6.1 Decision Rule of Proposed Method

In the proposed method shown in figure 6.1, both the ECG and the speech signals will be assumed to be recorded by the acquisition system that has both microphone and instrumentation amplifier. While in the acquisition process, filters will be applied to both signals to eliminate baseline wander, muscle noise, powerline noise, or unwanted frequency components. It is thought that the minimum recording time for both the training and testing stage will be 10 seconds. Originally, the system designated to require 10 seconds speech signal from each user by requesting of reading given text which differs from each other contains all the words in the alphabet. However, making it easy for other researchers to compare the results, we changed it to the publicly available database. In our experiments, time constrain is not set for the training process. Half of ECG and speech signals were separated into a test folder, whereas the other half were train folders. Then both ECG and speech signals were passed through a pre-processing process where inconsistent beats and spikes removed from the ECG database, whereas the silence part removed from the speech database. After the pre-processing stage, MFCC features were extracted from speech signals, whereas P-QRS-T features from ECG signals. Then vector quantization algorithm was applied onto P-QRS-T features in the train folder, and 16 significant P-QRS-T features were found. For the speech signal in the train and test folders, every MFCC features which

found in 10 seconds pass through vector quantization, and 32 significant MFCC features were extracted. By using the features in the train folder, two 1-D CNN classifiers were trained.

From now on, features from the test folder were used and given to the CNN algorithm for probabilistic scoring, where the higher the value of score indicates a higher chance of class ID, was it. The number of scores related to the number of classes in which we trained. If there is not a class which we can use to reject unauthorized user, the proposed system appoints all the feature to classes within known (genuine) dataset. For this reason, in the proposed system, class ID 0 was defined as an imposter class ID where unauthorized users were assigned to this class. If the value of the "ECG Threshold" is higher than zero, it means that we add a class where the unauthorized users were rejected for ECG based identification algorithm. It is the same with speech-based identification algorithm whenever the value of "Speech Threshold" is higher than zero. This newly appointed class is called the universal background model (UBM), and we can also call it Imposter Rejection Class (Class ID 0). In the training stage, ECG UBM and Speech UBM were also passed through the same process with genuine classes until the vector quantization. For the ECG UBM model, we found 1 significant P-QRS-T feature for each person's ECG signal in the UBM database, which consists of 152 people. For the Speech UBM model, we found 32 significant MFCC features for every person's speech signal, and a total of 4320 significant MFCC features were found for 135 people. Then, both 1-D CNN was trained by using features of both genuine classes and additional UBM class (Class ID 0).

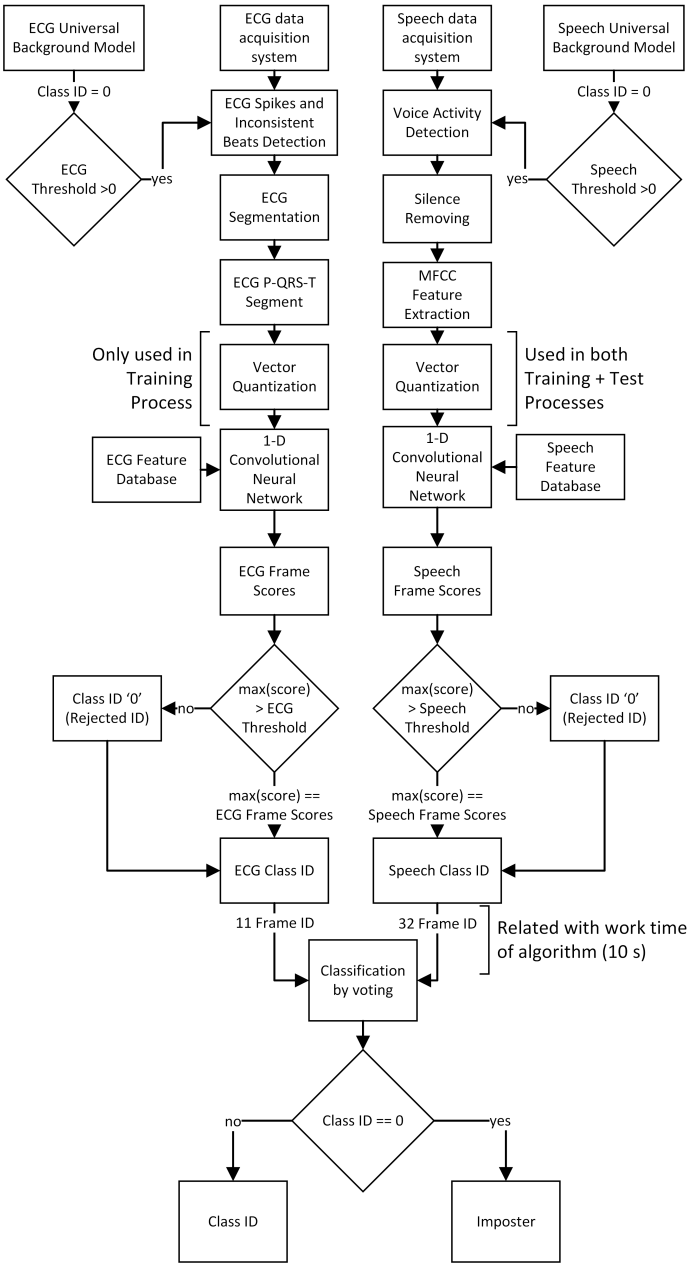


Figure 6.1: Decision Rule of the proposed algorithm

In the evaluation of the proposed system, two main experiments have been done where the threshold value is zero (which means no imposter rejection) or a threshold value where the intersection between False Acceptance Rate (FAR, imposter) and False Rejection Rate (FRR, genuine). If the maximum value in the scores is not higher than the threshold value, class ID was given as '0', which means reject the feature. If it is in another way, the class ID is given, which has a higher

score. In a given 10 seconds time, we accepted that every person has 11 P-QRS-T features whereas 1428 MFCC features. Every 1428 MFCC feature has passed through vector quantization, 32 significant MFCC features were extracted and given to the system with 11 P-QRS-T features. The class ID found from these features is then pass through the voting algorithm, where the most repetitive ID was set of the classifier's outcome. If the outcome is '0', then reject it from the system, which indicates that given features come from an imposter. If it is in another way, accept and give the ID of the most repetitive ones obtained from CNN classifiers.

The proposed system works like a traditional identification system if the value of "ECG threshold" and "Speech threshold" was given as zero, and there are multiple numbers of classes that contain only authorized users. In another way around, the proposed system works like a traditional verification system if the value of "ECG threshold" and "Speech threshold" are bigger than zero, and there are one genuine and multiple imposter classes giving to the system for testing it. Unlike the traditional verification and identification systems, our system exhibits both of their properties and rejects unauthorized users from the system where has multiple (genuine) classes in the system.

The important point is, speech signals in the train and test folders are passed through vector quantization whereas only ECG signals in the train folders were passed through vector quantization because of not varies over time. The equation below shows how to obtain the number of significant MFCC feature over the predefined working time of the system:

$$N = \frac{wt}{ft - (ft \;.\; ovl)} \tag{6.1}$$

,where "N" represents number of MFCC features found in a specific work time, wt represents work time of the algorithm (10 s in our application), ft represents MFCC framing time (0.02 s in our application), ovl represents the frame

overlapping percentage (0.65 in our application).

$$k = 2^{\lfloor \log_2 \frac{N}{cal} \rfloor} \tag{6.2}$$

,where *cal* is the calibration value (30 in our application), k is the number of significant MFCC features which we want to find by using vector quantization method.

6.2 Experimental Works

6.2.1 ECG Dataset

In our project, the combination of two fingertip datasets was used; the first one was the CYHBi database recorded by Hugo da Silva et al. [65] whereas the other one was a newly constructed database recorded by a self-made device [4]. In both databases, ECG signals were extracted from thumbs of people with a 1 kHz sampling rate. In the self-constructed database, records were taken from the 58 patients by two different weeks, having 1-5 minutes recording time. The dataset was created mostly taken from the students of Işık University which age differs between 19 to 26. CYHBi database, however, the 63 people's ECG signals were taken from two different months in 2 minutes recording time. The age of the CYHBi database differentiates between 20 to 24. The total number of fingertip databases became 121 by the combination of two datasets, but it decreased to 116 whenever the ECG spike and inconsistent beats algorithm was applied. 5 people's ECG signals were rejected to not having enough time or unsuitable for the recognition system where the movement of the patient was unstable. Fingertip ECG database was the main dataset for the evaluation of the system; however, for one of the following experiments, we will increase the dataset with ECG-ID database and remaining signals in the CYHBi database to test the performance of the proposed system when the number of the classes increases. For constructing the ECG-based Universal Background Model, we used 152 people

from the combination of PTB QT and MIT BIH Arrhythmia databases. Universal Background Model is crucial to reject unauthorized users when there is a small number of genuine classes.

6.2.2 Speech Dataset

In our project, two different datasets were used to test the performance. The first database was taken from LibriSpeech Corpus [66] where has a clean speech signal of 251 people and has 25 minutes recording time of each person. All the data in the database quantized with 16 bits, and the sampling rate was chosen as 16 kHz. We randomly selected the speech signals of 116 people and downsampled them into 8 kHz. The remaining speech signals of 135 people selected for constructing a universal background model.

The second database was taken from RedDots project [67] which was constructed to create challenges for the speech verification and identification system. In the RedDots database, there are speech signals of 62 people where some of the signals contain multiple noise sources such as the musical instrument sound in the background, mouse-clicking sounds of the user, the voice of crowds at the background, or has speakers whose pronouncing of the spoken language were bad.

6.2.3 Assessment Criteria

Four assessment criteria such as Accuracy, Sensitivity, Specificity, F_{score} were used to evaluate the system performance when the system has no imposter rejection feature. For the imposter rejection algorithm, four different criteria will be used, which are Imposter Accuracy, Genuine Accuracy, False Rejection Rate (Genuine), False Acceptance Rate (Imposter).

Sensitivity is defined as the proportion of positive classes that are correctly identified whereas, the specificity is defined as the proportion of the negative classes

that correctly identified. Equation for sensitivity and specificity are given as [68]

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6.3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6.4}$$

In the equation 6.3, 6.4 , TP, TN, FP, and FN are defined as True Positive, True Negative, False Positive, and False Negative, respectively. The definition for each can be given as

True Positive **(TP)** is defined as the outcome which the system correctly predicts the positive class.

True Negative **(TN)** is defined as the outcome which the system correctly predicts the negative class.

False Positive **(FP)** is defined as the outcome which the system incorrectly predicts the positive class.

False Negative **(FN)** is defined as the outcome which the system incorrectly predicts the negative class.

We take the average of TP, TN, FP, and FN for each class to evaluate Sensitivity, Specificity, and F-score because of having multiple classes.

F_{score} shows the accuracy of the experiment and defined as the harmonic mean of the precious and recall. The equation for F_{score} is given as [68]

$$F_{score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision + Recall}} \tag{6.5}$$

where Precision and Recall are calculated as the following equations

$$\text{Recall} = \text{Sensitivity} \tag{6.6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.7}$$

The other criteria is Accuracy and is defined as the ratio of number of classes correctly identified to the total number of attempts. It is calculated as the following equation [68]

$$\text{Accuracy} = \frac{N_{TP}}{N_F} \tag{6.8}$$

, where N_{TP} represents the number of correctly identified classes, and N_F represents the total number of attempts.

For the rejection algorithm, equation for finding the Imposter Accuracy, Genuine Accuracy, False Rejection Ratio (Genuine), False Acceptance Rate (Imposter) are given as

$$\text{Genuine Accuracy} = \frac{\text{Number of Genuine classes correctly identified}}{\text{Number of genuine attempts}} \tag{6.9}$$

$$\text{Imposter Accuracy} = \frac{\text{Number of Imposter classes correctly rejected}}{\text{Number of imposters attempts}} \tag{6.10}$$

$$\text{FAR} = \frac{\text{Number of incorrectly } \textbf{accept} \text{ access attempts by unauthorized users}}{\text{Number of unauthorized users attempts}} \tag{6.11}$$

$$\text{FRR} = \frac{\text{Number of false } \textbf{rejection} \text{ over genuine users}}{\text{Number of genuine users attempts}} \tag{6.12}$$

Chapter 7

Results and Discussion

7.1 Experiment 1

In first experiment, we started the algorithm with speech and ECG signals of 30, 45, 60, 75, 90 people, respectively. The subjects are divided into 3 folds, wherein each fold, speech, and ECG data are arranged according to the following rule: In the first fold, ECG records in the first week or month of the subjects were used for the training stage, whereas records in the second week or month were used for testing stage. For speech signal, the first half was used training whereas the second half for the test. In the second fold, the testing dataset was switched to training data. In the third and final fold, ECG and speech records which are in the tested and trained folders, are combined and selected in random order for test and train purposes. In Table 7.1, the number of P-QRS-T and MFCC features in each fold was given related with the number of subjects. By selecting ECG and Speech (rejection) threshold as zero, the performance evaluation was conducted for each independent 1-D CNNs and their fusion, and the results are given in Table 7.2 and Table 7.3. In the performance evaluation, four statistical measures were used, which are accuracy, specificity, sensitivity, and F_{score}.

Table 7.1: Properties of test and train sets

		Test Set		Train Set	
Number of Folds	**Number of Subjects**	**Number of P-QRS-T**	**Number of significant MFCC**	**Number of P-QRS-T**	**Number of significant MFCC**
Fold-1	30	1841	27827	480	25984
Fold-1	45	2715	35424	720	39168
Fold-1	60	3868	48032	960	53120
Fold-1	75	5265	61280	1200	67008
Fold-1	90	6292	74144	1440	81024
Fold-2	90	5783	81024	1440	74144
Fold-3	90	6037	75302	1440	82278

Table 7.2: Average accuracy of 3-fold cross validation

	30 people	45 people	60 people	75 people	90 people
ECG	95.85	94.38	89.96	90.43	90.22
Speech	99.66	98.43	98.00	98.67	97.94
Fusion	100.0	100.0	100.0	99.83	99.92

Table 7.3: Classification performance of the proposed method for each fold

ECG based Identification System					
Number of Folds	**Number of Subjects**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Fscore (%)**
Fold-1	90	88.5	99.8	86.7	91.7
Fold-2	90	85.3	99.7	84.2	89.5
Fold-3	90	96.7	99.9	95.5	97.6

Speech based Identification System					
Number of Folds	**Number of Subjects**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Fscore (%)**
Fold-1	90	97.5	99.9	97.6	98.7
Fold-2	90	96.6	99.9	96.4	98.1
Fold-3	90	99.5	99.9	99.3	99.6

The Proposed Fusion based Identification System					
Number of Folds	**Number of Subjects**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Fscore (%)**
Fold-1	90	100	100	100	100
Fold-2	90	100	100	100	100
Fold-3	90	99.9	99.9	99.7	99.8

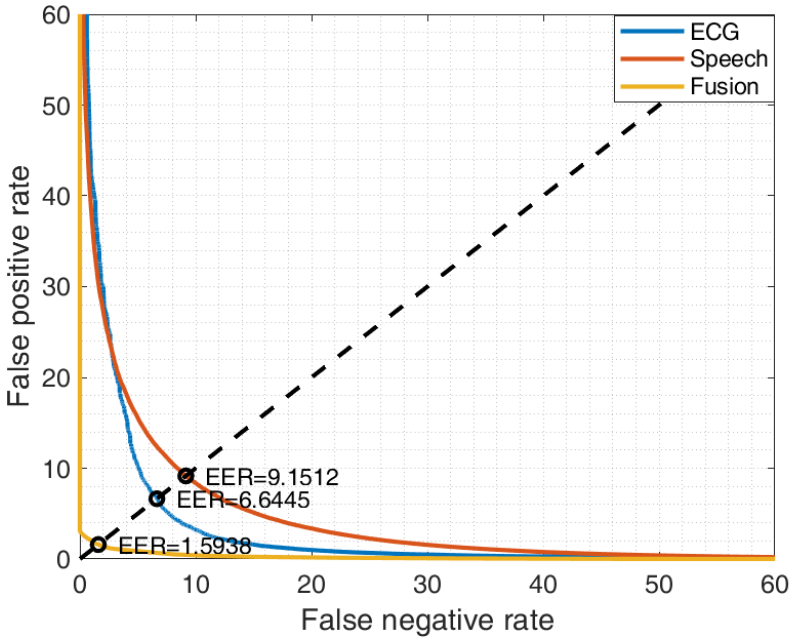


Figure 7.1: Detection Error Tradeoff

In the figure 7.1, the DET curve of the classifier for 90 people in the fold-1 can be seen. It was found by calculating the average score of each class. The equal error rate was found 9.2% for ECG-based system, 6.6% for the speech-based system, whereas 1.6% for the fusion of both systems.

7.2 Experiment 2

In second experiment, ECG and speech signal of 26 people set as imposter class, and 90 people were set as a genuine class. We know which class number was genuine, which class number was an imposter, so we trained the system with genuine 90 person's data. In the performance evaluation, speech and ECG (rejection) thresholds were calculated by finding the intersection of FRR (genuine) and FAR (imposter), where both of them were minimized (see in Fig 7.2, Fig 7.3). The ECG (rejection) threshold was found as 0.815, whereas 0.53 for the speech (rejection) threshold. Then, the performance of the system was evaluated by using the obtained rejection thresholds.

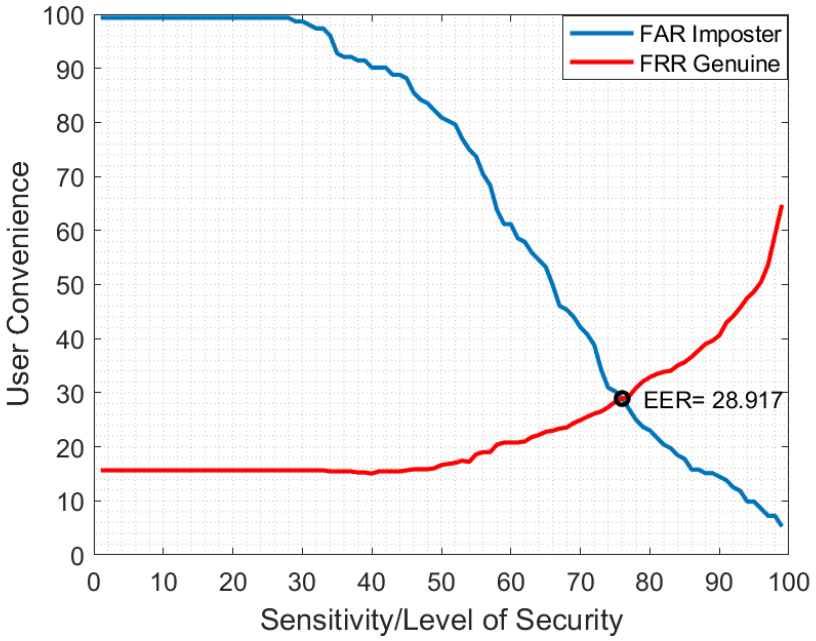


Figure 7.2: FRR vs FAR for ECG based Identification System

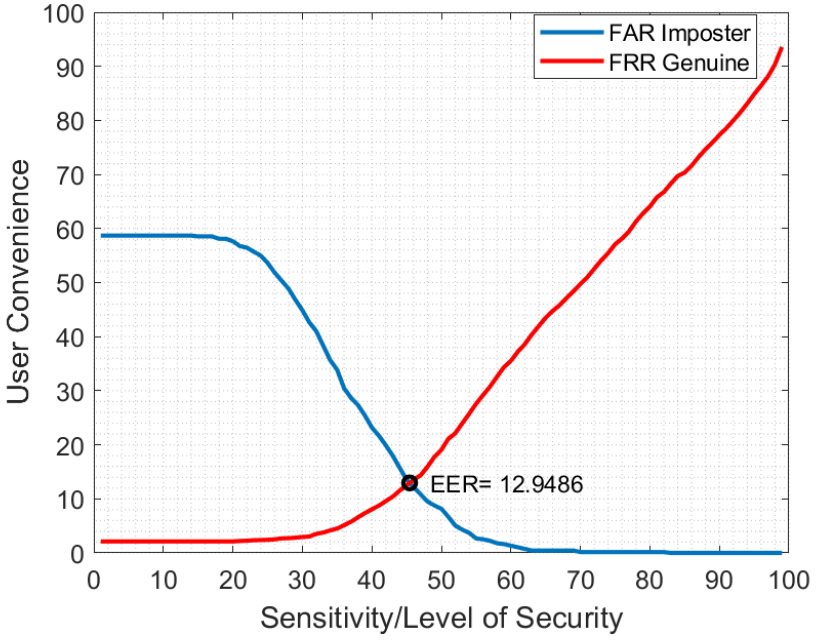


Figure 7.3: FRR vs FAR for Speech based Identification System

The increase of the rejection threshold provides better security, whereas decreasing the convenience, which makes users to do multiple attempts to be accepted

into the system and was correctly identified. It must be chosen wisely with respect to the application to be used.

Table 7.4: Genuine and Imposter Accuracy rates for 90 people

Proposed System	**Number of Subjects**	**Genuine Accuracy (%)**	**Imposter Accuracy (%)**	**Equal Error Rate (%)**
ECG	90	71.08	71.05	28.92
Speech	90	86.48	87.82	12.95
Fusion	90	91.68	96.05	6.82

In Figure 7.2, it is seen that system accepted all the imposter users when the value of "ECG threshold" are low. Increasing the value of the "ECG threshold" improves the imposter rejection accuracy whereas decreases the accuracy of correctly identifying genuine users. It is same for speech signal (see in figure 7.3); however, it differs of some point which is that; the increases of speech UBM dataset improves the imposter rejection accuracy significantly while the value of "Speech threshold" is still low. In our research, we increased the speech UBM dataset with 250 people and we saw that it decreases FAR (imposter) to 42% while "Speech threshold" is at "1" whereas it does not affect genuine accuracy. It indicates; increasing of speech UBM dataset significantly improves the imposter rejection performance. However, it can't be said for the ECG system because when we increased the ECG UBM dataset with 550 people, it only decreases the FAR (imposter) to 90% while "ECG threshold" is at "1". It also affected the accuracy of genuine users to decrease for the ECG system. For this reason, we suggest that rather than increasing the ECG UBM dataset, ECG UBM should be split into multiple (rejection) classes for not to affect the genuine accuracy rate.

Tables 7.5 and 7.6 show the accuracy rates of the system with respect to changes in batch size and learning rates of the CNN architecture, respectively. The CNN parameters were optimized according to our classification problem. The hyper-parameters used in the training of the CNN were tuned by comparing their result with each other. We concluded that the stochastic gradient descent with momentum (SGDM) was chosen as a better optimizer due to its learning speed and

Table 7.5: The performance of the proposed system tested with 90 genuine and 26 imposter people by changing the learning rates of the CNN (the batch size is 128)

	Genuine Accuracy (%)				Imposter Accuracy (%)			
Learning rate	**0.01**	0.005	0.001	0.0005	**0.01**	0.005	0.001	0.0005
ECG	**71.09**	66.07	70.50	72.87	**71.05**	76.32	73.03	73.03
Speech	**86.48**	87.57	85.66	86.18	**87.82**	88.71	87.82	87.07
Fusion	**91.68**	93.47	89.90	91.49	**96.05**	93.42	93.42	93.42

Table 7.6: The performance of the proposed system tested with 90 genuine and 26 imposter people by changing the batch size of the CNN (the learning rate is 0.01)

	Genuine Accuracy (%)				Imposter Accuracy (%)			
Batch size	64	**128**	256	512	64	**128**	256	512
ECG	70.10	**71.09**	76.04	73.66	71.05	**71.05**	75.66	73.68
Speech	84.14	**86.48**	87.70	88.17	85.14	**87.82**	88.11	87.37
Fusion	91.09	**91.68**	91.29	95.25	96.05	**96.05**	90.13	94.74

significant performance compared to the RMSProp or ADAM optimizers. ADAM and RMSProp learned our model in approximately 72 minutes, whereas SGDM learned it in 57 minutes. The optimum CNN learning rate and batch size for increasing the imposter rejection performance of the fusion system were found to be 0.01 and 128, respectively.

7.3 Experiment 3

In third experiment, we increased our ECG database by adding the remaining ECG signals measured at palm in the CYBHI database [65] and ECG signals measured at the arm in ECG-ID database [69] into our system. As a result, the ECG database was increased to 226 people where we randomly selected 176 people's data to be used as genuine classes, whereas 50 people for imposter classes. As for the speech database, we randomly selected 361 people in the "train-clean-360" dataset from LibriSpeech, which consists of 921 people. Then we randomly selected 176 people for genuine, 50 people for imposter classes, whereas the remaining 135 people for constructing a universal background model. We conducted

an experiment by using the previous (rejection) thresholds and compare the results with new threshold values found from the intersection of FRR and FAR. The result shows (see in table 7.7) that (rejection) threshold values changes whenever the dataset is increased, so suitable decision rule must be determined for balancing both imposter rejection and genuine acceptance accuracies.

Table 7.7: Genuine and Imposter Accuracy rates for 176 people

	Previous Rejection Thresholds		**New Rejection Thresholds**		
Proposed System	**Genuine Accuracy (%)**	**Imposter Accuracy (%)**	**Genuine Accuracy (%)**	**Imposter Accuracy (%)**	**Equal Error Rate (%)**
ECG	78.74	60.31	73.24	72.76	27.03
Speech	59.85	99.09	85.83	88.56	13.09
Fusion	74.61	100	89.54	96.10	10.3

7.4 Experiment 4

In our fourth experiment, the effect of the time constrain had been examined on 90 people's fingertip ECG and speech data in fold-1. We changed the working time of the system to 1, 3, 5, 10, 20, and 50 seconds, respectively, while using the databases in our previous experiments. If ECG or speech signals of the people in the database do not meet the required time constant, we evaluated the system by the maximum length of signal each person has. We conducted an experiment where each class has trained with sufficient features without time constraints. In performance evaluation, we tested the system by giving features to the system in a specific time range. In the end of experiment, accuracy of imposter rejection, and accuracy of genuine classes were given to be compared with each other (see in Tables 7.8 and 7.9).

Table 7.8: Identification performance for genuine people in a given response time

	1 sec	3 sec	5 sec	10 sec	20 sec	60 sec
ECG	62.07	64.62	66.75	71.08	71.98	72.00
Speech	55.09	71.38	79.28	86.48	89.95	90.34
Fusion	63.15	72.82	82.14	91.68	90.69	93.89

Table 7.9: Rejection performance for imposter people in a given response time

	1 sec	3 sec	5 sec	10 sec	20 sec	60 sec
ECG	64.22	66.78	68.56	71.05	72.72	75.00
Speech	55.44	70.71	79.98	87.81	89.29	90.09
Fusion	69.07	85.66	92.63	96.05	90.69	93.89

In Tables 7.8 and 7.9, the system's performance increases when we increase the response time of the system. In the Tables, it can also be seen that genuine and imposter accuracy rates differ from each other for each response time. This indicates that the exact intersection of FRR and FAR is not found when the signals are very hard to differentiate. Therefore, we used the thresholds where the accuracy of imposter rejection is higher than the accuracy of genuine identification. The genuine and imposter accuracy rates in the first 4th columns are to be expected the same such as the accuracy rate of their fusion in the 5th and 6th columns.

7.5 Experiment 5

In our fifth experiment, we exchanged LibriSpeech database with RedDots database [67] which has 62 speakers, contains 49 male and 13 female speakers from 21 countries. The records were taken through mobile crowd-sourcing, with the benefit of a potentially wider population and greater diversity. The language used in the database is English; however, the dataset was constructed to increase the difficulty of the identification system where has noises sources at the background or pronouncing of the spoken language were bad. The database was used to simulate the performance of the proposed algorithm in a bad conditional recorded speech signal. The results are given in Table 7.10.

Table 7.10: Genuine and Imposter Accuracy rates for RedDots speech database

Proposed System	**Number of Genuine Classes**	**Number of Imposter Classes**	**Genuine Accuracy (%)**	**Imposter Accuracy (%)**	**Equal Error Rate (%)**
ECG	48	14	75.51	77.64	23.93
Speech	48	14	67.10	67.34	32.87
Fusion	48	14	73.52	85.36	26.05

7.6 Experiment 6

In our final experiment, we randomly selected the number of 1, 3, 5 genuine classes and the remaining classes in the 116 people left for imposter classes. Then we trained the proposed system with genuine classes. In addition to the genuine class number, we exchanged person/people in the genuine class/classes three times from the database. Then we conducted a test to evaluate the identification accuracy (genuine) and rejection accuracy (imposter) of the proposed system, and the results are given in Table 7.11

Table 7.11: Accuracy Rates of the proposed system when few people were registered in the system

Average of Randomly Selected 3 Group of People (Genuine Class)				
Proposed System	**Number of Genuine Classes**	**Number of Imposter Classes**	**Genuine Accuracy (%)**	**Imposter Accuracy (%)**
ECG	1	115	100.00	90.67
ECG	3	113	81.37	87.07
ECG	5	111	70.93	70.93
Speech	1	115	100.00	99.60
Speech	3	113	99.40	98.47
Speech	5	111	98.53	98.37
Fusion	1	115	100.00	99.93
Fusion	3	113	100.00	98.93
Fusion	5	111	100.00	98.63

7.7 Discussion and Comparison of Relevant Researches

It must be known that there is no significant development over Speech and ECG fusion-based identification system. For this reason, we can not compare the proposed fusion-based system performance. However, we can compare the system

performance by taking into consideration of modalities that we used in our system.

In Tables 7.12 and 7.13, Recent research has been shown by comparing them with respect to the database, electrode replacement, or methodology, the number of subjects, identification rate, and working time. It can be seen that most of the research has not exceeded the number of 100 subjects for both ECG and speech identification systems. For speech identification systems, our proposed system has superior to most of the research in comparison of identification rate and working time. For the ECG-based identification system, our proposed system achieved a moderate identification rate by comparing the overall system. However, it should be considered that the proposed ECG-based system works with fingertip ECG signals, and it gives the result of the identification in 10 seconds. Most of the research ignores the system response time. So we think it is superior to other systems with respect to response time. The other superiority of our system is that it can work with both verification, identification, and authentication. If the number of user in the system is limited to 1 person, it works like verification system where reject the unauthorized people. If there are multiple users in the database, it will identify the person with no unauthorized rejection option or with an unauthorized rejection option that can be selectable by the users' will.

Table 7.12: Comparison of the Speech based System with Recent Researches

Author	Year	Database	Method	NS		Results(%)	WT
S. Chakroborty et al. [29]	2007	Polycost	Text-Indep.	131	IDR	81.6	754 UT
		YOHO		138	IDR	97.7	96 UT
Shung-Yung Lung [30]	2009	TALUNG	Text-Indep.	100	IDR	96	∼5 s
		KING		51	IDR	98	NA
Zhanyu Ma et al. [33]	2011	TIMIT	Text-Indep.	25	IDR	99.5	3 STC
Khaled Daqrouq [34]	2011	SC	Text-Indep.	29	IDR	91.1	4 STC (120s)
Hesham Tolda [35]	2011	SC	Text-Indep.	10	IDR	80	1 STC
A.D. Jafeer et al. [36]	2012	Switchboard	Text-Indep.	40	IDR	100	6 UT (∼20s)
S.S. Nidhyananthan et al. [37]	2013	SC	Text-Indep.	120	IDR	80.4	∼40
Jalil Taghia et al. [47]	2013	TIMIT	Text-Indep.	100	IDR	93.57	∼9 s
Hong Yu et al. [39]	2014	TIMIT	Text-Indep.	20	IDR	97.6	∼9 s
N. Almaadeed et al. [40]	2014	Grid	Text-Indep.	34	IDR	97.5	NA
N. M. AboElenein et al. [41]	2016	CHAINS	Text-Indep.	36	IDR	91	2 s
N. Almaadeed et al. [42]	2016	YOHO	Text-Indep.	NA	IDR	94.23	3 s
		NIST			IDR	92.15	
		TI_digits1			IDR	96.87	
		TI_digits2			IDR	97.34	
Zhanyu Ma et al. [52]	2017	TIMIT	Text-Indep.	100	IDR	99.52	3 UT
A. S. Imran [44]	2019	MOOC	Text-Indep.	119	IDR	93.37	3 s
					IDR	94.44	5 s
					IDR	94.64	7 s
Chao Zhang et al. [70]	2019	LibriSpeech	Text-Indep.	251	IDR	99.08	4 s
Qian-Bei Hong et al. [71]	2019	LibriSpeech	Text-Indep.	50	IDR	82.99	17 s
				50	EER	5.35	
				1172	IDR	73.26	
				1172	EER	6.89	
The proposed Speech Identification System(No Imposter Rejection)	**2020**	**LibriSpeech**	**Text-Indep.**	**90**	**IDR**	**97.93**	**10s**

where **UT** refers to Utterance, **STC** refers to sentence

Table 7.13: Comparison of the ECG based System with Recent Researches

Author	Year	DB	ER	NS		Results(%)	WT
Biel at al. [72]	2001	SC	Chest	20	IDR	98	NA
Shen et al. [73]	2002	MIT-BIH	Chest	20	IDR	100	∼1 s
A. Lourenco et al. [74]	2011	SC	Fingertip ECG	16	IDR	94.3	∼25 s(30 HB)
Z. Zhao and L. Yang [5]	2011	QT	Chest	20	IDR	95.3	NA
Shen et al. [6]	2011	SC	Palm	168	IDR	95.3	NA
Sara Zokaee et al. [7]	2012	SC	NA	50	IDR	94.7	NA
Fufu Zeng et al. [8]	2012	MIT-BIH Arr.	Chest	37	IDR	95.8	NA
Emna Rabhi et al. [9]	2013	MIT-BIH	Chest	18	IDR	99	10 min
A.C. Matos et al. [10]	2013	SC	L/R Index Finger	10	IDR	100	30 s
J. Wu and Y. Zhang [11]	2014	MIT-BIH Arr.	Chest	33	IDR	97.1	10 s
Kuo-Kun Tseng et al. [12]	2014	MIT-BIH NSR	Chest	18	IDR	95.3	NA
Huan Zhang et al. [14]	2015	MIT-BIH	Chest	36	IDR	94.4	NA
Juan Sebastian et al. [15]	2015	SC	Fingertip ECG	10	IDR FAR	81.82 1.41	4 s
M. Dai et al. [17]	2015	SC	Hand	14	IDR	77.15	NA
G. Altan et al. [18]	2015	ECG-ID	Wrists	90	IDR	91.5	20 s
M. N. Dar et al. [19]	2015	ECG-ID	Wrists	47	IDR	95.9	20 s
Xiafei Lei et al. [20]	2016	PTB	Chest + Limbs	100	IDR	99.3	NA
Kuo-Kun Tseng et al. [21]	2016	MIT-BIH NSR	Chest	18	IDR FAR	91.7 0	NA
M. Bassiouni et al. [22]	2016	MIT-BIH Arr.	Chest	30	IDR	97	NA
L. Wieclaw et al. [23]	2017	SC	Fingertip ECG	18	IDR	96	NA
Gang Zheng et al. [24]	2017	SC	NA	28	IDR	98.1	NA
Ronald Salloum et al. [25]	2017	ECG-ID	Wrists	18	IDR	100	∼8 s
M. Bassiouni et al. [26]	2018	MIT-BIH Arr. ECG-ID	Chest Wrists	30 90	IDR IDR	100 99	20000 s ∼127 s
Hakan Gürkan et al. [75]	2019	MIT-BIH Arr.	Chest	46	IDR	99.3	NA
N. Samarin [76]	2019	SC	Fingertip ECG	48	EER	9.7	∼96 s
Jae-Neung Lee et al. [27]	2019	SC	Arm	95	IDR	98.25	NA
The proposed ECG Identification System (No Imposter Rejection)	**2020**	**SC + CYBHi**	**Fingertip ECG**	**90**	**IDR**	**90.22**	**10s**

where **DB** refers to Database, **ER** refers to Electrode Replacement, **NS** refers to number of subjects, **IDR** refers to identification rate, **EER** refers to equal error rate, **FAR** refers to false acceptance rate, **WT** refers to work time which indicates the ECG frame in time while evaluating the system, **SC** refers to self-constructed, **NA** indicates that information is not available or computable, **HB** refers to Heart beat.

Chapter 8

Conclusion

We presented a fusion-based system which has both identification and verification feature that can reject the imposter classes. The proposed fusion algorithm was developed working on real-time security applications where there are multiple users. In the algorithm, we provided a solution for the degradation of fingertip ECG signals caused by the subject's movement. The proposed fusion method works with the principle of the voting method by looking at the outcome of each independent CNN system. The first experimental result shows that the proposed fusion-based system achieved a 100% accuracy rate for 90 people when there is no imposter rejection feature. The second experimental result shows that our algorithm rejected the imposter classes with a 91.68% accuracy rate whereas accepted with a 96.05% accuracy rate for 90 people. In the third experiment, we increased both the speech and ECG database and evaluated the performance of the system with 176 genuine, 50 imposter classes, and we achieved 89.54% accuracy rate in the genuine classes whereas 96.1% imposter rejection accuracy. In the fourth experiment, we compared both genuine and imposter class accuracy rates by changing the working time of the system. In the fifth experiment, the speech dataset exchanged with the RedDots dataset, given 48 genuine classes and 14 imposter classes, the proposed method has achieved 73.52% genuine identification accuracy whereas 85.36% imposter rejection accuracy. In the final experiment, the proposed system has been shown that it can also work when a few people are registered.

References

[1] A. K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology 14 (1) (2004) 4–20.

[2] S. Chauhan, A. Arora, A. Kaul, A survey of emerging biometric modalities, Procedia CS 2 (2010) 213–218. doi:10.1016/j.procs.2010.11.027.

[3] S.-C. Fang, H.-L. Chan, Human identification by quantifying similarity and dissimilarity in electrocardiogram phase space, Pattern Recognition 42 (2009) 1824–1831. doi:10.1016/j.patcog.2008.11.020.

[4] G. Guven, H. Gürkan, U. Guz, Biometric identification using fingertip electrocardiogram signals, Signal, Image and Video Processing 12 (2018) 933–940.

[5] Z. Zhao, L. Yang, ECG identification based on matching pursuit, Vol. 2, 2011, pp. 721–724. doi:10.1109/BMEI.2011.6098470.

[6] D. Shen, W. Tompkins, Y. H. Hu, Implementation of a one-lead ECG human identification system on a normal population, Journal of Engineering and Computer Innovations 2 (2011) 12–21.

[7] S. Zokaee, K. Faez, Human identification based on ECG and palmprint, International Journal of Electrical and Computer Engineering (IJECE) 2. doi:10.11591/ijece.v2i2.292.

[8] F. Zeng, H.-N. Huang, S.-Y. Tu, J.-S. Pan, A new statistical-based algorithm for ECG identification, 2012, pp. 301–304. doi:10.1109/IIH-MSP.2012.79.

[9] E. Rabhi, Z. Lachiri, Biometric personal identification system using the ECG signal, Vol. 40, 2013, pp. 507–510.

[10] A. Matos, A. Lourenco, J. Nascimento, Embedded system for individual recognition based on ECG biometrics, Procedia Technology 17. doi:10.1016/j.protcy.2014.10.236.

[11] J.-J. Wu, Y. Zhang, ECG identification based on neural networks, 2014, pp. 92–96. doi:10.1109/ICCWAMTIP.2014.7073368.

[12] K.-K. Tseng, J. Luo, R. Hegarty, W. Wang, D. Haiting, Sparse matrix for ECG identification with two-lead features, The Scientific World Journal (2015) 656807doi:10.1155/2015/656807.

[13] R. Horowitz, J. Morganroth, C. Parrotto, C. Chen, J. Soffer, F. Pauletto, Immediate diagnosis of acute myocardial infarction by two-dimensional echocardiography, 1982, p. 323–329. doi:10.1161/01.cir.65.2.323.

[14] H. Zhang, L. Chong, D. Guo, A fusion of ECG signal identification method, in: Proceedings of the 2015 International Conference on Education, Management, Information and Medicine, Atlantis Press, 2015/04, pp. 33–38. doi:https://doi.org/10.2991/emim-15.2015.7.

[15] J. Arteaga-Falconi, H. Al Osman, A. El Saddik, ECG authentication for mobile devices, IEEE Transactions on Instrumentation and Measurement 65 (2015) 1–10. doi:10.1109/TIM.2015.2503863.

[16] S. Israel, J. Irvine, A. Cheng, M. Wiederhold, B. Wiederhold, ECG to identify individuals, Pattern Recognition 38 (2005) 133–142. doi:10.1016/j.patcog.2004.05.014.

[17] M. Dai, B. Zhu, G. Zheng, Y. Wang, A method of ECG identification based on weighted correlation coefficient, in: Biometric Recognition, Springer International Publishing, Cham, 2015, pp. 633–640.

[18] G. Altan, Y. Kutlu, ECG based human identification using logspace grid analysis of second order difference plot, in: 2015 23nd Signal Processing and Communications Applications Conference (SIU), 2015, pp. 1288–1291.

[19] N. Dar, M. Akram, A. Shaukat, M. Khan, ECG based biometric identification for population with normal and cardiac anomalies using hybrid HRV and DWT features, 2015, pp. 1–5. doi:10.1109/ICITCS.2015.7292977.

[20] X. Lei, Y. Zhang, Z. Lu, Deep learning feature representation for electrocardiogram identification, in: 2016 IEEE International Conference on Digital Signal Processing (DSP), 2016, pp. 11–14.

[21] K. Tseng, D. Lee, W. Hurst, F. Lin, W. Ip, Frequency rank order statistic with unknown neural network for ECG identification system, in: 2016 4th International Conference on Enterprise Systems (ES), IEEE Computer Society, Los Alamitos, CA, USA, 2016, pp. 160–167. doi:10.1109/ES.2016.27.

[22] M. Bassiouni, A machine learning technique for person identification using ECG signals, IOSR Journal of Applied Physics 1 (2016) 37.

[23] L. Wieclaw, Y. Khoma, P. Fałat, D. Sabodashko, V. Herasymenko, Biometrie identification from raw ECG signal using deep learning techniques, in: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Vol. 1, 2017, pp. 129–133.

[24] G. Zheng, S. Ji, M. Dai, Y. Sun, ECG based identification by deep learning, 2017, pp. 503–510. doi:10.1007/978-3-319-69923-3_54.

[25] R. Salloum, C. . J. Kuo, ECG-based biometrics using recurrent neural networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2062–2066.

[26] M. Bassiouni, E.-S. El-Dahshan, W. Khalefa, A.-B. M.Salem, Intelligent hybrid approaches for human ECG signals identification, Signal, Image and Video Processing 12 (2018) 941–949. doi:10.1007/s11760-018-1237-5.

[27] J.-N. Lee, K.-C. Kwak, Personal identification using a robust eigen ECG network based on time-frequency representations of ECG signals, IEEE Access PP (2019) 1–1. doi:10.1109/ACCESS.2019.2904095.

[28] S. Marinov, I. För, S. Högskolan, I. Skövde, Text dependent and text independent speaker verification systems. technology and applications.

[29] S. Chakroborty, A. Roy, G. Saha, Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks, International Journal of Signal Processing 4 (2007) 114–122.

[30] S.-Y. Lung, Improved wavelet feature extraction using kernel analysis for text independent speaker recognition, Digital Signal Processing 20 (2010) 1400–1407. doi:10.1016/j.dsp.2009.12.004.

[31] S. Chakroborty, G. Saha, Feature selection using singular value decomposition and QR factorization with column pivoting for text-independent speaker identification, Speech Commun. 52 (9) (2010) 693–709. doi:10.1016/j.specom.2010.04.002.

[32] M. S. Sinith, A. Salim, K. Sankar, K. Narayanan, V. Soman, A novel method for text-independent speaker identification using MFCC and GMM, 2010, pp. 292 – 296. doi:10.1109/ICALIP.2010.5684389.

[33] Z. Ma, A. Leijon, Super-dirichlet mixture models using differential line spectral frequencies for text-independent speaker identification,interspeech, 2011, pp. 2349–2352.

[34] K. Daqrouq, Wavelet entropy and neural network for text-independent speaker identification, Eng. Appl. Artif. Intell. 24 (5) (2011) 796–802. doi:10.1016/j.engappai.2011.01.001.

[35] H. Tolba, A high-performance text-independent speaker identification of arabic speakers using a CHMM-based approach, alexandria engineering journal 50 (2011) 43–47.

[36] S. J. Abdallah, I. M. Osman, M. E. Mustafa, Text-independent speaker identification using hidden markov model, 2012.

[37] S. Nidhyananthan, R. Selva Kumari, Language and text-independent speaker identification system using GMM 9.

[38] J. Taghia, Z. Ma, A. Leijon, On von-mises fisher mixture model in text-independent speaker identification, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2013) 2499–2503.

[39] H. Yu, Z. Ma, M. Li, J. Guo, Histogram transform model using MFCC features for text-independent speaker identification, Conference Record - Asilomar Conference on Signals, Systems and Computers (2015) 500–504doi:10.1109/ACSSC.2014.7094494.

[40] N. Almaadeed, A. Aggoun, A. Amira, Speaker identification using multimodal neural networks and wavelet analysis, Biometrics 4. doi:10.1049/iet-bmt.2014.0011.

[41] N. Aboelenien, K. Amin, M. Ibrahim, M. M. Hadhoud, Improved text-independent speaker identification system for real time applications, in: 2016 Fourth International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), 2016, pp. 58–62. doi:10.1109/JEC-ECC.2016.7518967.

[42] N. Almaadeed, A. Aggoun, A. Amira, Text-independent speaker identification using vowel formants, J. Signal Process. Syst. 82 (3) (2016) 345–356. doi:10.1007/s11265-015-1005-5.

[43] L. R. Rabiner, B. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice Hall, 1993.

[44] A. Imran, Z. Kastrati, T. Svendsen, A. Kurti, Text-independent speaker ID employing 2D-CNN for automatic video lecture categorization in a MOOC setting, 2019, pp. 273–277. doi:10.1109/ICTAI.2019.00046.

[45] X. lu, J. Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification, Speech Communication 50 (2008) 312–322. doi:10.1016/j.specom.2007.10.005.

[46] B. Heard, Lecture slides, Atlantic Cape Community College CHAPTER 2013 Pearson Education.

[47] Springhouse, ECG strip ease: An arrhythmia interpretation workbook, 2012.

[48] G. Gargiulo, True unipolar ECG machine for wilson central terminal measurements, BioMed Research International 2015 (2015) 1–7. doi:10.1155/2015/586397.

[49] A. Fratini, M. Sansone, P. Bifulco, M. Cesarelli, Individual identification via electrocardiogram analysis, BioMedical Engineering OnLine 14 (1) (2015) 78. doi:10.1186/s12938-015-0072-y.

[50] Invensense, Microphone Array Beamforming Application Notes, 2015.

[51] J. Pan, W. J. Tompkins, A real-time QRS detection algorithm, IEEE Transactions on Biomedical Engineering BME-32 (1985) 230–236.

[52] Y. Zhang, Z. Xiong, J. Mao, L. Ou, The study of parallel k-means algorithm, Vol. 2, 2006, pp. 5868 – 5871. doi:10.1109/WCICA.2006.1714203.

[53] D. MacKay, "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

[54] D. S. Jat, A. S. Limbo, C. Singh, "Intelligent Speech Signal Processing" Academic Press, 2019.

[55] J. Sohn, S. Kim, W. Sung, A statistical model based voice activity detector, Signal Processing Letters, IEEE 6 (1999) 1 – 3. doi:10.1109/97.736233.

[56] S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, Acoustics, Speech and Signal Processing, IEEE Transactions On (1980) 357–366.

[57] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, Q. Tian, HMM-based audio keyword generation, in: Advances in Multimedia Information Processing - PCM 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 566–574.

[58] X. Huang, A. Acero, H.-W. Hon, Spoken language processing: A guide to theory, algorithm, and system development.

[59] J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, 2018, pp. 7–19. doi:10.1201/9780429500459.

[60] M. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015. URL https://books.google.com.tr/books?id=STDBswEACAAJ

[61] A. NG, Machine learning course Lecture on Supervised Learning. URL http://cs229.stanford.edu/materials.ht

[62] Convolutional neural network, MATLAB (R2016a).

[63] B. Ginsburg, Application case study—machine learning, 2017, pp. 345–367. doi:10.1016/B978-0-12-811986-0.00016-9.

[64] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. Inman, 1D convolutional neural networks and applications: A survey, ArXiv abs/1905.03554.

[65] H. Plácido da Silva, A. Lourenco, A. Fred, N. Raposo, M. Aires-de Sousa, Check your biosignals here: A new dataset for off-the-person ECG biometrics, Computer methods and programs in biomedicine 113. doi:10.1016/j.cmpb.2013.11.017.

[66] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, 2015, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

[67] K. A. Lee, A. Larcher, H. ARONOWITZ, G. WANG, P. Kenny, The RedDots challenge: Towards characterizing speakers from short utterances, interspeech 2016, san francisco, 2016.
URL https://sites.google.com/site/thereddotsproject/reddots-challenge

[68] H. Gurkan, U. Guz, S. Yarman, A novel biometric authentication approach using electrocardiogram signals, 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2013) 4259–4262doi:10.1109/EMBC.2013.6610486.

[69] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000 (June 13)) e215–e220, circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[70] C. Zhang, W. Chen, C. Xu, Depthwise separable convolutions for short utterance speaker identification, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2019, pp. 962–966.

[71] Q. Hong, C. Wu, M. Su, H. Wang, Sequential speaker embedding and transfer learning for text-independent speaker identification, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 827–832.

[72] L. Biel, O. Pettersson, L. Philipson, P. Wide, ECG analysis: a new approach in human identification, IEEE Transactions on Instrumentation and Measurement 50 (3) (2001) 808–812.

[73] T. W. Shen, W. J. Tompkins, Y. H. Hu, One-lead ECG for identity verification, in: Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology, Vol. 1, 2002, pp. 62–63 vol.1.

[74] A. Lourenco, H. Plácido da Silva, A. Fred, Unveiling the biometric potential of finger-based ECG signals, Computational intelligence and neuroscience (2011) 1–8doi:10.1155/2011/720971.

[75] H. Gürkan, H. Ayça, Evrişimsel sinir ağı ve QRS imgeleri kullanarak EKG tabanlı biyometrik tanıma yöntemi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 26 (2020) 318 – 327.

[76] N. Samarin, D. Sannella, A key to your heart: Biometric authentication based on ECG signals, ArXiv abs/1906.09181.