DEGREE PROGRAMME IN ELECTRICAL ENGINEERING or
DEGREE PROGRAMME IN WIRELESS COMMUNICATIONS ENGINEERING

SPEAKER GENDER RECOGNITION SYSTEM

Author	Zimeng Hong
Supervisor	Mika Ylianttila
Second Examiner	Madhusanka Liyanage
Technical Advisor	Jude Okwuibe

May, 2017

Zimeng H. (2017) Speaker Gender Recognition System. University of Oulu, Department of Communications Engineering. Master's Degree Programme in Wireless Communications Engineering. Master's Thesis, 54 p.

ABSTRACT

Automatic gender recognition through speech is one of the fundamental mechanisms in human-machine interaction. Typical application areas of this technology range from gender-targeted advertising to gender-specific IoT (Internet of Things) applications. It can also be used to narrow down the scope of investigations in crime scenarios.

There are many possible methods of recognizing the gender of a speaker. In machine learning applications, the first step is to acquire and convert the natural human voice into a form of machine understandable signal. Useful voice features then could be extracted and labelled with gender information so that are then trained by machines. After that, new input voice can be captured and processed and the machine is able to extract the features by pattern modelling.

In this thesis, a real-time speaker gender recognition system was designed within Matlab environment. This system could automatically identify the gender of a speaker by voice. The implementation work utilized voice processing and feature extraction techniques to deal with an input speech coming from a microphone or a recorded speech file. The response features are extracted and classified. Then the machine learning classification method (Naïve Bayes Classifier) is used to distinguish the gender features. The recognition result with gender information is then finally displayed.

The evaluation of the speaker gender recognition systems was done in an experiment with 40 participants (half male and half female) in a quite small room. The experiment recorded 400 speech samples by speakers from 16 countries in 17 languages. These 400 speech samples were tested by the gender recognition system and showed a considerably good performance, with only 29 errors of recognition (92.75% accuracy). In comparison with previous speaker gender recognition systems, most of them obtained the accuracy no more than 90% and only one obtained 100% accuracy with very limited testers. We can then conclude that the performance of the speaker gender recognition system designed in this thesis is reliable.

Key words: Gender Recognition, Speaker, Matlab, Naïve Bayes Classifier, Machine Learning.

TABLE OF CONTENTS

FOREWORD

This thesis is aimed at designing a speaker gender recognition system on Matlab platform. This work was done under Profesor Mika Ylianttila's research group, Centre for Wireless Communications (CWC), Department of Communications Engineering at the University of Oulu.

I would like to use this opportunity to express my appreciation and gratitude to my supervisor Prof. Mika Ylianttila for offering me this work under his guidance and support. To my second examiner, Dr. Madhusanka Liyanage, I am very appreciated for his inspiring advice to start this thesis. I would like to extend my appreciation to Jude Okwuibe for his meticulous guidance and help.

This thesis gives acknowledgement to the Naked Approach project, which is a significant strategic research project that investigates means to create an alternative, human-driven approach to operate in the digitalizing society. I am appreciated the funding support from Tekes: Finnish Funding Agency for Innovation and this project partners.

I would like to thank all the professors at CWC for their constructive suggestions. Additionally, I want to express my great appreciation to Prof. Kari Kärkkäinen with his advices and support from the first day I came to University of Oulu to the following two years.

Oulu, May 19th, 2017

Zimeng Hong

LIST OF ABBREVIATIONS AND SYMBOLS

A/D	Analog-to-Digital
ASR	Automatic Speech Recognition
CC	Cepstral Coefficient
DFT	Discrete Fourier Transform
ECOC	error-correcting output codes
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GUI	Graphical user interface
HCI	Human-computer interaction
HNR	Harmonics-to-Noise Ratio
HMM	Hidden Markov Model
IoT	Internet of Things
KNN	A nearest-neighbor
LPC	Linear Predictive Coding
MAP	Maximum a posterior
MFCC	Mel-frequency Cepstral Coefficient
MPM	McLeod Pitch Method
Bit	Binary digit
RASTA-PLP	Representation Relative Spectra-Perceptual Linear Predictive
SNR	Signal-to-Noise ratio
SVM	Support Vector Machines
YAAPT	Yet Another Algorithm for Pitch Tracking
ZCR	Zero-crossing Rate
C	Class variable
F_s	Sampling rate
F_{mean}	Mean voice fundamental frequency
F_{min}	Minimum voice fundamental frequency
F_{max}	Maximum voice fundamental frequency
f	Actual voice frequency
$f_{centroid}$	Voice frequency centroid
f_i	Frequency of spectrum at window i of n
f_{mean}	Mean voice frequency
f_{median}	Median voice frequency
f_{mode}	Mode voice frequency
$f_{kurtosis}$	Kurtosis of voice frequency
f_{sd}	Standard deviation of voice frequency
f_{sfm}	Spectral flatness of voice frequency
f_{spent}	Spectral entropy of voice frequency
$f_{skewness}$	Skewness of voice frequency
Hz	Hertz
j	Number of independent feature vector in x
k	Number of vector in class variable set C
$Mel(f)$	Perceived frequency
M	Number of samples

m	Number of independent feature vector in x
N	Total number of samples in each frame
n	Number of frequency windows in the spectrum
μ	Pre-emphasis filter coefficient
Q_{25}	First quartile
Q_{75}	Third quartile
Q_{IQR}	Interquartile range
X_k	Frequency magnitudes
x	Independent variables set
y_i	Relative amplitude of the frequency f_i
$\hat{y}$	A class label
Z	Scaling factor
z	Filter
argmax	Arguments of the maxima
E(t)	Represents the expected value of the quantity t
Σ	Summation operator
Π	Capital Pi notation
√	Square root operator
%	Percentage
ln(x)	Natural logarithm is the logarithm to the base e of a number
$P(A)$	Probability of A
$P(A\|B)$	Probability of A given B
$w(n)$	Window operation

1. INTRODUCTION

1.1. Motivation

Voice is one of the most common means of communication in the world. The vibration of an object called the sound source causes surrounding air molecules to vibrate and spread. Such continuous sound vibration in the air or other media is what gives rise to sound. Human ears acquire and perceive the voice signal depending on the frequency of the sound, which allows us to distinguish the voice from different speakers or sound sources. Generally, voice contains a large number of sound waves with different frequencies, by which humans can recognize the attributes of each individual voice.

In real life, human ears could usually verify a speaker by listening to the linguistic information. A voice contains a lot of linguistic information of the speaker. Human ears can identify some natural attributes (gender, age, origin, etc.) even when the voice comes from unfamiliar speakers. Therefore, voice features could be regarded as voiceprints capable of identifying the genders, ages, origins and emotional states of the speakers.

The speaker gender recognition system proposed in this work automatically extracts features from a speech signal and those features will be used to determine the gender of the speaker. This gender recognition system can be leveraged upon in some practical applications, such as determining the gender of the user which will be conducive to providing more targeted services based on gender interoperability. Also, the speaker gender recognition system applied in HCI (Human-computer interaction) could specify the user interface scope and improve the user experience in most IoT applications. Such information could be used to provide gender customizations in such IoT applications and promote the security level in such applications [1].

It is significant to recognize the gender of speaker automatically according to the speech signals. Firstly, the user gender recognition could help to improve the interactive function of information systems, so that it can automatically select the appropriate interaction service for different gender of users. With the advances in information technology and the overall progress of the society, human-machine technology has been deeply involved as part of our daily life. In human-machine interaction, the demand of a system service is different for both male and female users, such as user interface style and preferences of words and color. Therefore, the distinction of the gender for different users using such voice authentication scheme would certainly improve the overall efficiency of human-computer interaction. Secondly, we can make targeted access control through the determination of the gender of users. In social networking system. For instance, user gender identification can be used to control access to male visit zone or female visit zone.

Also, the global consumer product industry can benefit from such gender identification system. This comes as a natural benefit, given that different genders of consumers have their respective consumption tendencies for certain products, which will then narrow down the scope of advertising for enterprises save the corresponding costs. Gender recognition can also be applied in customer relationship management, according to telephone counseling to identify the gender of customers, and the marketing strategies will be effectively developed.

1.2. Background

The speaker gender recognition is the technology to determine the gender of a speaker by analyzing and processing a specific input speech signal. In 1988, Childers and Bae reported the gender recognition of the speakers and the speaker recognition has been developed as a specialized research filed [2]. At the early 1990s, the study of speaker gender identification focused on the selection and extraction of relative features such as Mel-frequency Cepstral Coefficient (MFCC), Pitch, Linear Predictive Coding (LPC) and voice fundamental frequency, which are used as the parameters for speaker recognition research at that moment [3]. In the late 1990s, Cepstral Coefficient (CC), prosodic feature, formant (F1, F2, F3) and other features were also applied in the field of speaker gender recognition and achieved better recognition results [4].

In the past 20 years, the speech gender recognition research based on recognizing word or syllable has been developed towards the use of continuous speech signal. Further development of speech recognition technology utilized acoustic features such as LPC and Representation Relative Spectra-Perceptual Linear Predictive (RASTA-PLP) in speaker gender recognition by voice [5]. In addition, with the successful application of Support Vector Machines (SVM) and Gaussian Mixture Model (GMM) in speech signal processing research, the accuracy of speaker gender recognition technology has been significantly improved. GMM could be regarded as a special situation of Hidden Markov Model (HMM), because of its convenient computation and good robustness, its widely work as recognition pattern in speaker recognition system. Additionally, Wavelet Analysis, Time-Frequency Analysis and Neural Network also applied in the development of speaker recognition.

The research in gender recognition application is becoming popular. In the quiet environment, the gender recognition accuracy by speech can reach an accuracy of 90% [6]. The work of Abdulla, W. H., Kasabov, N. K., and Zealand, D. N. was tested with TIMIT continuous speech corpus, and KEL isolates words speech corpus and showed 100% gender discrimination accuracy (no error recorded) [7]. Ali, M. S., Islam, M. S., and Hossain, M. A. in 2012 designed a system with a gender recognition accuracy of 80% [8]. Sheikh, H. designed a system in 2013 providing almost 80% accuracy on 10 speakers [9]. Erokyar, H. in 2014 designed an age and gender recognition system for speech application, and obtained average 64.2% with 108 experiment speakers.

1.3. The structure of thesis

This thesis aimed at designing a speaker gender recognition system by processing speech signal based on Matlab platform with the assumption that input speech signal is obtained under a quiet and small-scale environment. This designed speaker gender recognition modelling (shown in Figure 1) indicates input speech signal after three general steps (pre-processing, feature extraction and classification), and this modelling will finally give a gender result which is either male or female.

Figure 1.General speaker gender Recognition Modelling.

The structure of this thesis is organized as follows. Chapter 2 is a literature review about pre-processing, voice feature extraction method and Naïve Bayes classifier. This chapter introduces characteristic voice features that can be extracted from a typical voice sample such as fundamental frequency and classifications based on Naïve Bayes classifier. Chapter 3 described the implementation work of this designed. The implementation consists of voice recording, pre-processing, feature extraction and classification in backend, and a graphical user interface to output the tested results with recording function in frontend. Chapter 4 was the experimental work for evaluating the performance of this recognition system. There were 40 (half male and half female) volunteers from 16 countries participating in this experiment, and totally 400 speech samples were recorded in 17 languages. Chapter 5 gave a further discussion of this designed speaker gender recognition system in the experimental point of view. Additionally, a performance comparison of different classification applied in this designed system was discussed and also the future work was illustrated in Chapter 5. A summary of this thesis was concluded in Chapter 6.

2. LITERATURE REVIEW

The designed system firstly requires a voice pre-processing technique, which deals with the natural human speech from analog formation into the form of machine understandable digital signal. Then pre-processing technique will remove the silent parts as well as the noise voice using End-point detection, and voice signal can be enhanced by pre-emphasis technique. After this, processed speech is demonstrated as time-domain signal, and might be transform into frequency-domain through Fast Fourier Transform. In this designed system, voice features are extracted from frequency-domain speech signal computed by the algorithm into voice frequency related features and voice fundamental frequency related features. Finally, the machine learning classification method Naïve Bayes classifier is applied for the purpose of recognizing the pattern of a voice feature.

In this chapter, literature review of voice pre-processing, voice feature extraction and voice feature extraction was described. The first section introduced voice pre-processing knowledge including Analog-to-Digital (A/D) conversion, pre-emphasis, frame blocking and hamming window, Fast Fourier Transform, and End-point detection. The following section illustrated the detail algorithm of voice feature extractions where 15 voice features were extracted from voice frequency and voice fundamental frequency for the designed speaker gender system. The last section introduced the classification which was used to identify the gender from a speaker.

2.1. Voice Pre-Processing

During the speech transmission, it is unavoidable for noise interference and voice attenuation. After the natural voice obtained by the micro-speaker handset, it was then formatted as analog signal whereas the digital form of voice signal which can be analyzed easier. The collected voice signal data exists in the time domain, and the conversion of time domain signal into frequency domain signal is indispensable. Therefore, the primary step after obtaining speech though micro-speaker handset is voice processing, and it is also known as voice pre-processing.

This section introduced voice pre-processing techniques. The first subsection presented A/D conversion technique. After that, pre-emphasis technique was explained by an equation. In the third subsection, frame blocking and hamming window were introduced. The fourth section mainly referred to a time-to-frequency domain conversion technique (Fast Fourier Transform). The last subsection described end-point detection including noise and silence signal removal technique.

2.1.1. A/D Conversion

The speech signal is a continuous analog signal variation in both time axis and amplitude axis. The first step is to handle speech signal to obtain discrete digital voice signal by sampling, quantization and A/D conversion.

A micro-speaker handset can be used to receive the acoustic wave as an analog signal as shown in Figure 2. This analog signal is conditioned with antialiasing filtering which limits the bandwidth of the signal to approximately the Nyquist rate (half the sampling rate). According to the Nyquist sampling law [10], the sampling

rate should be more than twice of the original signal rate so that the information would not get lost during sampling process and the original signal can be reconstructed accurately from the sampling signal. Since the usable voice frequency band ranges of normal human speech are approximately from 300 Hz to 3400 Hz [11], the sampling rate is one of the factors limiting the voice quality. This can be achieved by restricting the frequency response (the highest audio signal that can be carried) to one-half of the sampling rate.

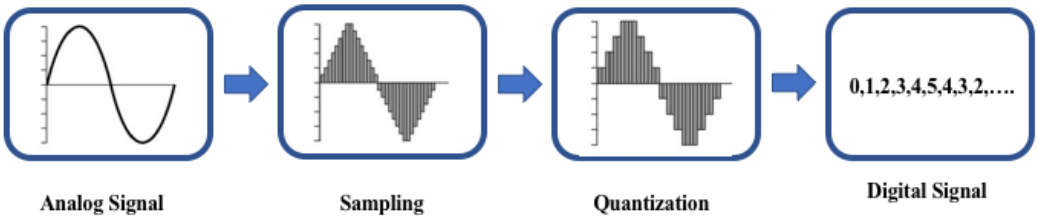

Figure 2.The process of Analog signal converts to Digital signal.

Figure 2 also indicated that after voice signal sampling, quantization of voice signal will be done before the digital signal formed from conditioned analog signal. Finally, the acoustic sound pressure wave is transformed into digital signals.

2.1.2. Pre-emphasis

There is attenuation at high-frequency segments since the voice signal speaks out from the mouth. Therefore, enhance the voice signal by using pre-emphasis technique before voice feature extraction is necessary.

The pre-emphasis filter was used to flatten the speech waveforms and glottal waveforms spectrally. The purpose of pre-emphasis is filtering low frequency interference, especially power frequency interference at low frequency segment, and enhancing the high frequency portion of voice recognition to make the spectrum of the voice signal flatter in order to produce a high-pass filter to carry out spectral analysis. Voice signal pre-emphasis has commonly happened after A/D conversion by the first-order digital pre-emphasis filter equation [12] is

$$H(z) = 1 - \mu z^{-1}, \quad (1)$$

where z represents filter, μ is pre-emphasis filter coefficient with the value ranging commonly between 0.9~1.

2.1.3. Frame Blocking and Hamming window

Frame blocking refers to the process of segmenting the speech signal obtained after pre-emphasis filtering into the number of N small frames, with adjacent frames separated by M $(M < N)$. The first frame consists of the first N samples and the second frame start M samples after the first frame, therefore, the first overlaps will be started at the $(N - M)$ samples. Then the third frame will begin at the position of the

M sample after the second frame with an overlap of $(N - 2M)$ samples, the rest frames can be done with the same approach. In general, the values of N and M are 256 and 100 for the numbers of the frames and samples, respectively.

The procedure of hamming window is carried out after frame blocking. Each frame of the windows is aimed at minimizing speech signal discontinuities before and after each frame. The analytical representation of hamming window is given by

$$w(n) = 0.54 - 4.46\cos(\frac{2\pi n}{N-1}) \qquad 0 \leq n \leq N-1, \tag{2}$$

where $w(n)$ is the window operation, n is the number of each individual sample and N is the total number of the speech samples [12, 13].

This method is popularly used in MFCC before mel frequency warping step (as introduced Section 2.2.3) where mel scales was calculated.

2.1.4. Fast Fourier Transform

Fast Fourier Transform (FFT) algorithm is used for computing the Discrete Fourier Transform (DFT) of a sequence, or its inverse form. In the voice signal processing, FFT converts each frame of those N samples from time domain signal into the formulation of frequency domain [13].

The FFT is a computationally efficient implementation of the DFT method, which is defined on the set of N samples $\{x_n\}$ as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}},\ k = 0,1,2,\dots,N-1. \tag{3}$$

In general, X_k is a complex number and we only consider its absolute value (frequency magnitudes) here. The resulting sequence $\{x_k\}$ can be interpreted as follows: the positive frequencies $0 \leq f < \frac{1}{2}F_s$ correspond to the values $0 \leq n \leq \frac{1}{2}N - 1$, while the negative frequencies $-\frac{1}{2}F_s < f < 0$ correspond to $\frac{1}{2}N + 1 \leq n \leq N - 1$. Here, F_s represents the sampling rate. The obtained result after this step is often referred to as frequency spectrum of the voice signal.

2.1.5. End-point Detection

End-point detection technique detects the starting point and ending point of a given speech signal so that the pure voice signal and noise signal can be distinguished. An effective end-point detection can not only reduce the amount of data collection in voice recognition system and save the processing time, but also eliminate the interference from noise or silence to improve the performance of voice recognition system [14].

The difference between pure voice signal and noise signal is the energy difference. The energy of voice signal is larger than that of the noise signal. When the signal-to-noise ratio (SNR) is up to certain level, the calculation of short time energy or short-time average amplitude of input signal will become possible based on

short-time energy end-point detection method and the distinguishing of the pure voice segment and noise segment can be realized.

Another detection method for the separation of pure voice segment and unvoiced segment can be realized by Zero-crossing Rate (ZCR), which analyzes continuous voice signal in time domain. ZCR counts the number of times that a speech signal passed through the horizontal time axis, that is, the number of signal samples which change the number of symbols.

ZCR is suitable to detect pure voice segment while the short-time energy to detect unvoiced segment. In the beginning, two thresholds were designed for short-time energy and ZCR. The lower threshold (value A) which is sensitive to the variation of a speech signal is easy to exceed, and the other threshold (value B) should be larger than the lower one. When the signal value is over the low threshold A, it does not definitely indicate the starting point of a speech signal. It might be also caused by the noise signal. A signal segment could be specified as a real voice signal only when it meets the following two characteristics: (1) the signal value exceeds the lower threshold A for a certain period, and (2) it is higher than the threshold B during this period.

Therefore, End-point detection method can be applied to divide a speech signal as silence segments and pure voice segments.

2.2. Voice Features Extraction

Variations of pronunciation habits result in voice uniqueness based on the attributes of gender, emotion, age, etc. There are many acoustic characteristics affect the output of a voice signal from a vocal tract. Jitter and shimmer are quantified as the cycle-to-cycle fundamental frequency and amplitude of a sound waveform, both of which can influence the stability of a speech signal [15]. Harmonics-to-Noise Ratio (HNR) represents the energy of the periodic harmonic compared to the energy of noise, quantifying the amount of additive noise in the voice signal [16]. Articulation rate is a prosodic feature described as the speech speed of a speaker, and the variations of this rate can usually be caused by the diversity of individual vocal organs. Duration of speech pause means total length of suspension interval within the speech signal per unit time. The number of speech pauses represent the total number of pause in speech signal per unit. In the speaker recognition system, accurate extraction of the voice features is of great help for efficient voice signal processing.

This section introduced voice features. Features related to voice frequency spectrum are discussed firstly. The voice fundamental frequency was described in the following subsection. Mel-frequency Cepstral Coefficients was introduced in the last subsection.

2.2.1. Voice Frequency Relevant Features

A voice frequency or voice band is one of the frequencies, within part of the audio range, which is used for the transmission of speech. Since voice signal is transferred from a time domain into a frequency domain, many characteristics of frequency are acquired as voice features. The voice frequency spectra can be calculated as twelve voice frequency related features, including mean frequency, standard deviation, median frequency, first quartile, third quartile, interquartile range, skewness, kurtosis, spectral entropy, spectral flatness, mode frequency, frequency centroid. This section

explained the definition of the features related to these twelve voice frequency spectra and also the necessary equations.

Mean Frequency

Mean frequency of a speech, denoted as f_{mean}, could be deemed as the average value of speech frequency within a specific frequency spectrum range. It is calculated as the sum of the frequency of the spectrum divided by the total number of frequency frames in a spectrum. The equation for the calculation of mean frequency is

$$f_{mean} = \frac{\sum_{i=0}^{n} f_i}{n}, \tag{4}$$

where f_{mean} is the mean frequency of a voice, n is the number of frequency frames in the spectrum, f_i is the frequency of spectrum at frame i of n.

Standard Deviation

The standard deviation of speech frequency, represented as f_{sd}, could be an extent to the frequency frames in a spectrum differing from the mean frequency. The standard deviation value of a featured voice frequency spectrum is a measure how each frequency frames is clustered from the mean frequency [17]. The standard deviation of a specific speech frequency equation is defined as

$$f_{sd} = \sqrt{\frac{\sum_{i=0}^{n}(f_i^2 - f_{mean}^2)}{n}}, \tag{5}$$

where f_{sd} denotes the value of the standard deviation of a speech voice, f_{mean} is aforementioned mean frequency value, n refers to the number of frequency frames in the spectrum, and f_i is the frequency of spectrum at frame i of n.

Median Frequency

The median frequency of speech, represented as f_{median}, could be understood as the median value of speech frequency in a specific frequency spectrum range. The value of median frequency is represented as f_{median}, which can be calculated by the intensity of the signal in the whole spectrum, and divided by two with equal amplitude. Then median frequency is selected at which the cumulative intensity is defined as

$$f_{median} = \frac{\sum_{i=0}^{n} f_i}{2}, \tag{6}$$

where f_{median} denotes the median frequency value of a speech, f_i is the frequency of spectrum at frame i of n.

First Quartile

The first quartile of voice frequency, represented as Q_{25}, is one of the voice frequency features denoted by Q_{25}, which is the median of the lower half of the

speech frequency set. It means that about 25% of the frequency values in the speech frequency set lie below Q_{25} and 75% of speech frequency values larger than Q_{25} [18].

Third Quartile

The third quartile of voice frequency, represented as Q_{75}, is one of voice frequency features denoted by Q_{75}, which is the median of the upper half of the speech frequency set. It suggests that about 75% of the frequency values in the speech frequency set lie below Q_{75} and only 25% of speech frequency values lie above Q_{75} [18].

Interquartile Range

The interquartile range of voice frequency is one of the voice frequency features denoted as Q_{IQR}, which means the frequency ranging between Q_{25} and Q_{75}. The value of Q_{IQR} is obtained by the difference between Q_{75} and Q_{25} of a specific speech.

Skewness

Skewness, represented as $f_{skewness}$, is a measure of the asymmetry of the voice frequency spectrum around the sample mean. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero. If skewness is negative, the spectrum spreads out more to the left of the mean than to the right side, and vice versa. The skewness of speech frequency spectrum[1] is defined as

$$f_{skewness} = \frac{E(\sum_{i=0}^{n} f_i - f_{mean})^3}{f_{sd}^3}, \tag{7}$$

where $f_{skewness}$ denotes the skewness value of a voice frequency, f_{mean} is the mean frequency value, f_i is the frequency value of spectrum at window i of n, f_{sd} is the standard deviation value, $E(t)$ represents the expected value of the quantity t.

Kurtosis

Kurtosis, represented as $f_{kurtosis}$, is a measure of how outlier-prone a distribution is. When kurtosis value equals to 3, the frequency spectrum will exhibit a normal shape. When distributions that are less outlier-prone have a kurtosis value less than 3, the frequency spectrum will have fewer items at the center and the tails than the normal curve but with more items in the shoulders. When distributions that are more outlier-prone than the normal distribution have a kurtosis value greater than 3, the frequency spectrum will have more items near the center and at the tails, with fewer items in the shoulders compared to a normal distribution with the same mean and variance.[2] The kurtosis of a distribution[3] is defined as

$$f_{kurtosis} = \frac{E(f - f_{mean})^4}{f_{sd}^4}, \tag{8}$$

---

[1] The MathWorks. (2016). *Skewness.*
[2] Spectral properties, R Document, *seewave* version 2.0.4.
[3] The MathWorks. (2016). *Kurtosis*

where $f_{kurtosis}$ denotes the kurtosis value of voice frequency, f_{mean} is the mean of frequency value, f_{sd} is the standard deviation value, $E(t)$ represents the expected value of the quantity t.

Spectral Entropy

The spectral entropy of voice frequency, represented as f_{spent}, is calculated by using Shannon entropy properly normalized and applied to the power spectrum density of the speech signal [19]. The equation of spectral entropy is

$$f_{spent} = \frac{-sum(\sum_{i=0}^{n} y_i * \log(\sum_{i=0}^{n} y_i))}{\log(n)}, \quad (9)$$

where f_{spent} denotes spectral entropy value of voice frequency, n is the number of frequency windows in the spectrum, y_i is the relative amplitude value of the frequency at window i of n and the total sum of y_i equals 1.

Spectral Flatness

The voice frequency feature spectral flatness, represented as f_{sfm}, is defined as the ratio between of geometric mean value to arithmetic mean value.[4]

Mode Frequency

The mode frequency of a voice frequency feature, represented as f_{mode}, refers to the most frequently occurring frequency value in a speech frequency spectrum dataset.[5]

Frequency Centroid

Frequency centroid, represented by $f_{centroid}$, refers to the centroid value in a speech frequency spectrum dataset and can be computed as

$$f_{centroid} = sum(f_i * y_i), \quad (10)$$

where $f_{centroid}$ denotes the frequency centroid value, f_i is the frequency of spectrum at window i of n, and y_i is the relative amplitude value of frequency.

2.2.2. Voice Fundamental Frequency Relevant Features

Voice fundamental frequency, also known as voice pitch, is applied to detect the duration of pitch time, which records the period from the opening of vocal folds to its ending. When people is speaking, airflow passes through the glottis to make a relaxation oscillation of vocal folds, and the generated quasiperiodic signal will excite vocal tract pronunciation and finally produce sound signal. The frequency of

---

[4] Spectral flatness measure, R Document, *seewave* version 2.0.4.

[5] The MathWorks. (2016). Most frequent values in array.

vocal folds is called voice fundamental frequency, with a corresponding pure tone pitch.

Voice pitch has been extensively used in speech compression encoding, speech synthesis, speech recognition and other related analysis aspects. Therefore, accurate pitch extraction and reliable fundamental frequency estimation are crucial to the speech signal processing. It has a dominant impact on the possibility of synthesized speech reconstruction of the original voice signal and the recognition rate of speech recognition.

In this section, voice fundamental frequency detection method and voice fundamental frequency related features were introduced respectively.

Voice Fundamental Frequency Detection

Voice fundamental frequency detection is significant for drawing the completely consistent pitch curve which tracks the changeable fundamental frequency of vocal folds. A voice fundamental frequency detection algorithm is designed to estimate the fundamental frequency of the speech signal, which can be generally realized in two domains: time domain and frequency domain.

In the time domain, the pitch detection algorithm estimates the period of quasiperiodic signal and then inverts the value in order to get the pitch or fundamental frequency. The simplest approach is to find the ZCR of the voice signal, which is a measurement of how often the waveform crosses zero per unit time [20]. However, the detection of ZCR might be difficult when the waveform is made by multiple quasiperiodic signals which usually bring a complex variation. The more complicated approach is utilizing autocorrelation between two waveforms to compare the dissimilarity at the time interval. Autocorrelation algorithm could provide accurate results of highly periodic signals without noisy signals or polyphonic sounds environment. The most accurate approach is premium autocorrelation algorithms, such as the YIN algorithm [21] and the McLeod Pitch Method (MPM) algorithm [22].

In the frequency domain, the signal is converted into the polyphonic frequency spectrum signals, therefore are possible to be detected. The popular converting method is FFT, which is suitable for achieving a good efficiency. The frequency-domain algorithm aims at matching the frequency characteristics with pre-defined frequency maps and detecting signal peaks. The most popular algorithms include the harmonic product spectrum [23], cepstral analysis [24], and maximum likelihood.

Additionally, spectral pitch detection algorithm, Yet Another Algorithm for Pitch Tracking (YAAPT) [25] algorithm combined both time domain processing by autocorrelation function and frequency domain processing utilizing spectral information identification. This approach reduces the tracking errors caused by time domain or frequency domain independent processing in another domain.

Voice Fundamental Frequency Features

At the technical level, a typical female adult speaks with a higher voice fundamental frequency than that of a male adult speaker. The voice fundamental frequency of a male is from 85Hz to 180Hz, while that of a female is from 165Hz to 225Hz [26]. Thus, the voice fundamental frequency of most speeches falls below the bottom of the ‘voice frequency’ band. However, enough of the harmonic series will be presented for the missing fundamental to create the impression of hearing the fundamental tone.

Voice fundamental frequency related features include mean voice fundamental frequency (F_{mean}), minimum voice fundamental frequency (F_{min}) and maximum voice fundamental frequency (F_{max}), which can be calculated by the average value, minimum value and maximum value of voice fundamental frequency respectively.

2.2.3. Mel-frequency Cepstral Coefficient

Mel-frequency Cepstrum representation of the short-term power spectrum of voice, is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [27]. Mel-frequency Cepstral Coefficient (MFCC) is a feature that widely used in voice signal recognition, especially in the automatic speech recognition (ASR), whose coefficients collectively make up a Mel-frequency Cepstrum.

MFCC Extraction

The block diagram for the computation of MFCC is shown in Figure 3. We can obtain MFCCs after speech signal as input signal after the following processing steps: framing, windowing, FFT, Mel frequency warping, and cepstrum.

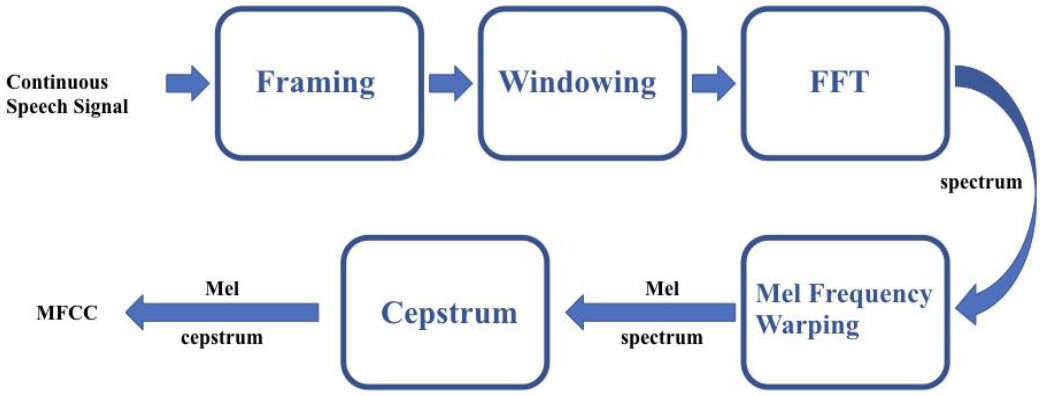

Figure 3. MFCC Extraction

Speech signal constantly changed, and the first step of MFCC detection is to split the speech signal into small frames in order to get relatively stable speech signal within each frame. If the size of the frame is shorter, the samples will be not enough for the estimation of a reliable spectrum. On the other hand, if the size of the frame is longer, the signal will change too much throughout the frame. In general, a speech signal is framing into 20-40ms frames. The next step in the processing is to window each individual frame as described in Section 2.1.3.

Secondly, the calculation of periodogram for each frame can be estimated by the power spectrum FFT (introduced in Section 2.1.4), to identify frequencies presented in the frame. Depending on location in the cochlea of the vibrations, different neural excitations inform the brain of certain frequencies. When the frequencies are far away

from the perceptible frequency range of human cochlea, the frequency acceptable capacity of human ear might decrease.

The next step is to apply the Mel filter-bank to the power spectra, and summarize the energy in each filter to check how much energy is present in each frequency region by avoidance of two tightly spaced frequencies. The first filter is very narrow and gives an indication of how much energy is present around 0 Hz. As frequencies get higher, the filters become wider and wider so that changes are less concerned.

Since human ear cannot hear loudness on a linear scale, it is necessary to amplify the perceived as much volume into voice. The step at this stage is taking the logarithm of all filter-bank energies, and the channel normalization technique allows usage of cepstral mean subtraction. Finally, compute the FFT of the log filter-bank energies. The FFT decorrelate the filter-bank energies because filter-banks are all overlapped and their energies are correlated to each other.

The Mel scale related to perceived frequency is measured frequency, which tells us exactly how to space our filter banks and how wide they should be made. The features will match more closely to what humans hear by incorporating such scale. For each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale [13]. The formula for converting frequency to Mel scale is:

$$Mel(f) = 1125 \times \ln(1 + \frac{f}{700}), \tag{11}$$

where f denotes the real frequency, and $Mel(f)$ denotes the perceived mel frequency.

In the final step, the log mel spectrum is converted back to the time-domain. The output is known as the MFCC. Since the mel spectrum coefficients are natural numbers, we are using Discrete Cosine Transform (DCT). $\widetilde{S_q}$, where $q = 1, 2, \dots Q$, denotes mel power spectrum coefficients than the MFCC's, and $\widetilde{C_p}$ can be calculated as follows:

$$\widetilde{C_p} = \sum_{p=1}^{P} (\log \widetilde{S_q}) \cos[p(q - \tfrac{1}{2}) \tfrac{\pi}{Q}]. \qquad q = 1, 2, \dots Q \tag{12}$$

We excluded the first component, $\widetilde{C_0}$, from the DCT since it represents the mean value of the input signal which carries little speaker specific information.

2.3. Classification

The problem of speaker gender recognition, which is a broader topic in scientific and engineering fields, is also called pattern recognition. It is aimed at classifying objects into one of classes or labels which are speaker genders in this case. In speaker gender pattern recognition, the extracted voice features (also called patterns here) comprise the training set and can be used to derive a classification algorithm. Since the classification procedure in speaker gender pattern recognition modelling is capable of extracting voice features, it is also mentioned as feature matching in many cases.

Classification is a type of machine learning, which means an algorithm designed to 'learn' the classification of new observations from examples of labeled data. This algorithm is a typical approach to extract features from an input speech, which makes it possible to predict a class or label. Before the algorithm predictor working, the

classifier needs to be trained with a dataset containing feature data with corresponding classes or labels. This step is necessary to input a training set into this algorithm.

The state-of-art classification feature matching includes three distinct algorithms: the statistic algorithm, bioinformatics algorithm and neutral network algorithm. Statistic classification algorithm includes discriminant analysis which assumed different generated data based on different Gaussian distributions[6]; logistic regression offers a nice probabilistic framework which can easily adjust the classification thresholds; Naïve Bayes classifier is designed in the case when predictors are independent of one another within each class. Decision trees or classification trees predict responses to data following the decisions in the tree from the root (beginning) node down to a leaf node[7]. Bioinformatics classification algorithm, for instance, Support Vector Machines provide greater accuracy and kernel-function choices on low-through medium-dimensional data sets. Neural network algorithm is popularly applied in deep learning by performing transformation learning with a pre-trained deep network.

This section mainly introduced Naïve Bayes classification in the first subsection accompanied with a demonstrated example in the second subsections.

2.3.1. Naïve Bayes Classification

The Naïve Bayes Classifier technique is based on the Bayesian theorem with strong independence assumptions between the features. Naïve Bayes is a simple technique for constructing the classifier that assign feature parameters drawn from the training set to the class labels. Naïve Bayes classifier assumes that all the value of features is independent of each other, in other words, there is no possible correlations between any two classes of features. Naïve Bayes classifier is highly scalable and worked quite well in complicated situation. It is especially suitable in the case when there are high inputs of the dimensionality features, requiring large number of linear feature parameters.

Naïve Bayes is a conditional probability model, which can classify a problem instance with m independent feature vectors represented by $= \{x_1, \dots, x_m\}$. The x values are assigned by instance probabilities $P(C_k|x_1, \dots, x_m)$ where each of k possible outcomes C_k [28]. Applying Bayes' theorem, the conditional probability can be decomposed as

$$P(C_k|x) = \frac{P(C_k)P(x|C_k)}{P(x)}, \tag{13}$$

It can be written with the Bayesian probability terminology,

$$posterior = \frac{Prior \times likelihood}{evidence}. \tag{14}$$

---

[6] The MathWorks. (2016). Discriminant analysis.

[7] The MathWorks. (2016). Statistics and Machine Learning Toolbox.

Using the naïve independence assumption

$$P(x_i|C_k, x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_m) = P(\big(x_j\big|C_m), \tag{15}$$

For all j values, the relationship in equation (13) is simplified as

$$P(C_k|\, x_1, \dots, x_m) = \frac{P(C_k)\prod_{j=1}^{m} P(x_j|C_k)}{P(\, x_1,\dots,x_m)}. \tag{16}$$

Since $P(\, x_1, \dots, x_m)$ is a constant input, the joint model can be expressed as

$$\begin{aligned} P(C_k|\, x_1, \dots, x_m) &\propto P(C_k, x_1, \dots, x_m) \\ &\propto P(C_k)P(x_1|\, C_k)P(x_2|\, C_k)P(x_3|\, C_k) \dots \\ &\propto P(C_k)\prod_{j=1}^{m} P\big(x_j \mid C_k\big) \end{aligned} \tag{17}$$

This means that under the independence assumptions in equation (13), the conditional distribution over the class variable C is

$$P(C_k|\, x_1, \dots, x_m) = \frac{1}{Z}\, P(C_k)\prod_{j=1}^{m} P\big(x_j \mid C_k\big), \tag{18}$$

where the evidence $Z = P(x)$ is a scaling factor merely dependent on $x_1, \dots, x_m$, that is, a constant if the values of the feature variables are known.

The construction of Naïve Bayes classifier can be derived from the independent feature probability model. Maximum a posteriori (MAP) decision rule can be used to estimate $P(C_k)$ and $P\big(x_j \mid C_k\big)$, which therefore assigns a class label $\hat{y} = C_k$ for k in the training set as

$$\hat{y} = \operatorname{argmax}_{k\in\{1,\dots,K\}} P(C_k)\prod_{j=1}^{m} P\big(x_j \mid C_k\big). \tag{19}$$

2.3.2. Demonstration of Naïve Bayes Classifier

To demonstrate the concept of Naïve Bayes Classifier, the following example displayed in Figure 3 indicates that the objects can be classified as either GREEN or RED. The task is to classify color species in new cases as they arrive. It is based on the currently exiting objects so that it can decide which class label they belong. There are twice as many GREEN objects as RED, therefore, it is reasonable to predict a non-observed new case is twice likely belongs to GREEN rather than RED. This prediction in the Bayesian analysis is known as the prior probability[8].

[8] Quest Software Inc. (2017). Naïve Bayes Classifier.

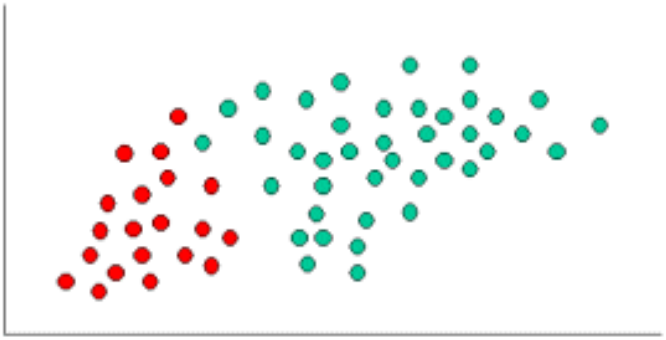

Figure 4.Exiting objects classified as GREEN or RED

Prior probabilities are based on previous experience, in this case, the percentage of GREEN and RED objects, and are usually used to predict the outcomes before they happen. In this case, assume there is a total of 60 objects, 40 of which are GREEN and 20 RED, the prior probability for class membership of GREEN and RED is written as:

$$Prior\ probability\ of\ GREEN \propto \frac{number\ of\ GREEN\ objects}{Total\ number\ of\ objects} = \frac{40}{60},$$

$$Prior\ probability\ of\ RED \propto \frac{number\ of\ RED\ objects}{Total\ number\ of\ objects} = \frac{20}{60}.$$

After formulated our prior probability, we are now ready to classify a new object, the WHITE circle as shown in Figure 2. Figure 2 indicated that both GREEN and RED objects are well clustered, therefore, it is reasonable to assume that more GREEN objects in the vicinity of WHITE object so that the WHITE object is more likely belong to GREEN color. For measuring the likelihood of WHITE object, Naïve Bayes classifier draws a circle around WHITE which encompasses a number (to be chosen a priori) of points irrespective of their class labels.

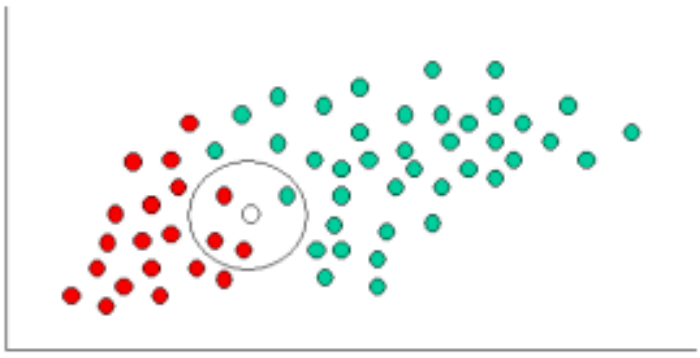

Figure 5.A circle is drawn around WHITE

Thus, we calculate the total number of points in the circle belonging to each class label and calculate the likelihood of WHTIE given both GREEN and RED possibilities:

$$Likelihood\ of\ WHITE\ given\ GREEN \propto \frac{number\ of\ GREEN\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ GREEN\ cases} = \frac{1}{4},$$

$$Likelihood\ of\ WHITE\ given\ RED \propto \frac{number\ of\ RED\ in\ the\ vicinity\ of\ X}{Total\ number\ of\ RED\ cases} = \frac{3}{4}.$$

From the illustration above, it is clearly shown that the likelihood of WHITE given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$Probability\ of\ WHITE\ given\ GREEN \propto \frac{1}{40},$$
$$Probability\ of\ WHITE\ given\ RED \propto \frac{3}{20}.$$

Overall, we can see that the prior probabilities indicate that WITHE may belong to GREEN, but the likelihood indicates the class membership of WHITE is RED. However, in the Bayesian analysis, the final classification is made by combing both prior probability value and likelihood value in order to form a posterior probability by Bayes' rule. The posterior probability of WHITE being GREEN and RED is:

$$Posterior\ probability\ of\ WHITE\ being\ GREEN \propto$$
$$Prior\ probability\ of\ GREEN\ \times\ Likelihood\ of\ WHITE\ given\ GREEN$$
$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60},$$

$$Posterior\ probability\ of\ WHITE\ being\ RED \propto$$
$$Prior\ probability\ of\ RED\ \times\ Likelihood\ of\ WHITE\ given\ RED$$
$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}.$$

Finally, we classify WHITE object as RED since its class membership achieves the largest posterior probability.

3. SYSTEM DESIGN AND IMPLEMENTATION

In this chapter, the practical implementation work of speaker gender recognition system was introduced. The practical implementation work was done by macOs Sierra (10.12.4) MacBook Air (13-inch, Early2014), 1.4 GHz Intel Core i5 processor, RAM is 4 GB 1600 MHz DDR3. To begin with, the implementation platform Matlab R2017a was simply introduced in Section 3.1. Then, the actual implementation procedure work in the backend of this speaker recognition system was described in detail. This designed system first provides voice recording function as Section 3.2.1 demonstrate, then the voice is recorded into Matlab. After voice pre-processing as Section 3.2.2 indicated, 15 voice features used in this designed system is extracted as Section 3.2.3 introduced. The finally recognition decision is made by Naïve Bayes classification as Section 3.2.4 described.

In additional to that, Section 3.3 demonstrates the public voice corpus used as the training database of classification in this designed system. Finally, a graphical user interface work in the frontend of this speaker recognition system was demonstrated.

3.1. Implementation Platform

The speaker recognition system can be implemented in many programming platforms. In this thesis, The Matlab was chosen as the main and the only platform.

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. With a proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, a creation of user interfaces, and interfacing with programs written in other languages. An optional toolbox allows an access to additional computing abilities, such as Simulink, adds graphical multi-domain simulation and model-based design for dynamic and embedded systems. Thus, Matlab has a widely scope of applications including signal and image processing, telecommunications, control system design, testing and measurement, financial modelling and analysis, computational biology and many other applications.

In this speaker recognition system implementation work, most coding work was done by the editor. Additionally, signal processing toolbox provides functions for voice pre-processing procedure such as resampling, removing silence, FFT analysis and so on. Statistics and machine learning toolbox provide classification function and app of classification learning used to measure the accuracy of a specific classifier with a set of training data. Matlab GUI provides point-and-click control of software application so that this speaker gender recognition systems act as an application with a controllable interface. Audio recording and playback program provides a real-time function so that this speaker gender recognition system could record voice and work out immediately.

3.2. Implementation of Speaker recognition system

This section introduced the entire implementation of this work, and the structure of the speaker recognition system was shown in Figure 6. Human natural speech as an input signal was first recorded by Matlab and converted into digital form, and the

detailed information was introduced in the following Section 3.2.1. Afterwards, Section 3.2.2 described that the digital form of speech signal coming into pre-processing step, which processed the speech signal and transformed the time-domain signal into the frequency-domain signal. The voice frequency as well as its fundamental frequency were consequently obtained. The next step was to extract the features based on voice frequency and voice fundamental frequency which were represented in Section 3.2.3. After that, classification procedure was demonstrated in Section 3.2.4. In the last part, the training set from public resources was described.

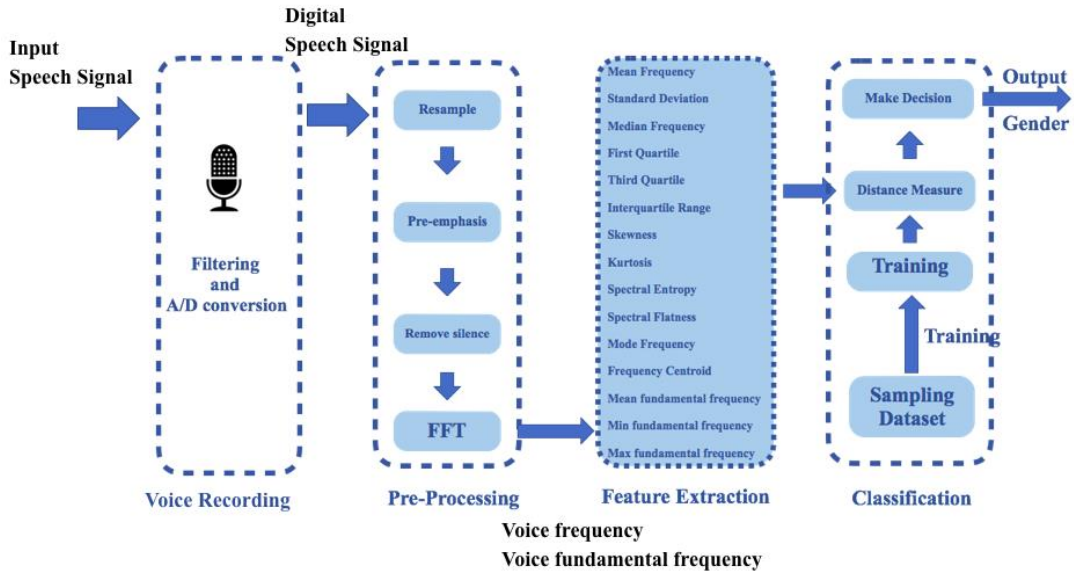

Figure 6. Structure of speaker recognition system

3.2.1. Voice Recording

Speech was recorded into Matlab directly with a speaker handset, which is actually an earphone microphone in this experiment. Matlab function performs voice recording (as shown in Figure 6) commanded as '*audiorecorder*' to create an audio object. In this designed speaker recognition system, default speech sampling rate is 11025 Hz (a sampling rate over 800 Hz can already promise a good quality) which means 11025 samples per second with 16 bits per sample. A mono channel type was selected by default. It means that only one voice channel is needed during recording.

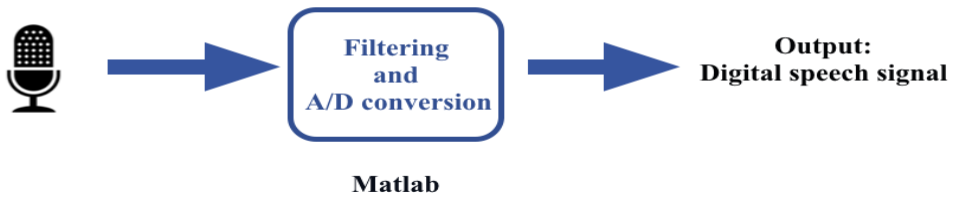

Figure 7. Voice recording by Matlab

The audio file was converted in accordance with the format of audio data matrix in time domain through the Matlab function commanded '*audioread*', and then a m-by-1 matrix could be returned, where m is the number of audio samples recorded in the file. The *'sound'* function was adopted to convert speech matrix data to sound play.

3.2.2. Pre-processing

The digital speech signal matrix was processed by Pre-Processing before the feature extraction. The first step was resampling the time-domain speech matrix into decimation or interpolation for the purpose of returning to the original sampling rate of 11025 Hz, as shown in Figure 8. Then the signal was input into the pre-emphasis filter, which was designed according to the equation in Section 2.1.2. The pre-emphasis filter coefficient μ here was fixed as 0.95, which means that 95% of each sample is presumably originated from the previous sample.

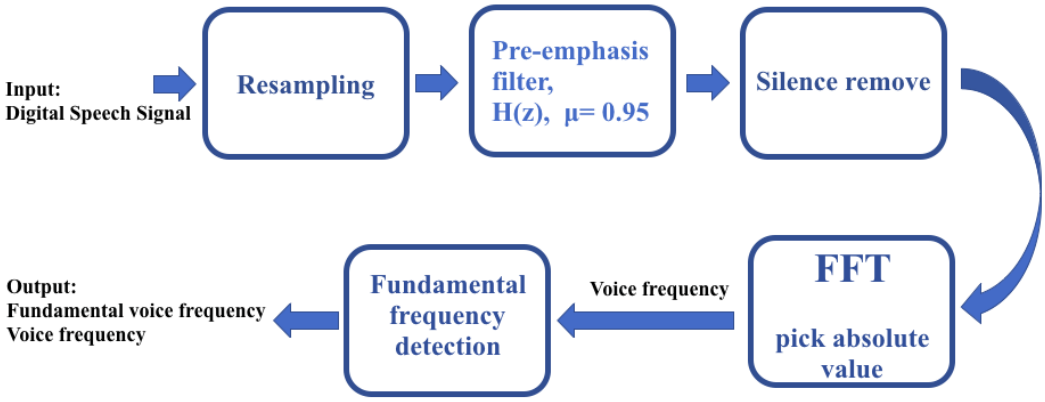

Figure 8. Demonstration of pre-processing

Since the value of silence data in the speech signal matrix is 0, the next step was to remove all the '0' data in order to get rid of all the silent part. Time-domain signal was converted into the frequency-domain signal by fast Fourier transform as Section 2.1.4 presented, which could be realized by the *'fft'* command within Matlab platform [29]. However, the obtained frequency-domain data may also contain negative values, however, only absolute values are needed in the processing. The fundamental frequency could be acquired by a function getF0[9]. In addition, both voice frequency and voice fundamental frequency were limited between 20Hz to 280Hz, which is assumed as the human acoustic range.

[9] getF0 (2006), a Matlab function created by Raul Fernandez.

As a result, input digital speech signal matrix after pre-processing was processed and supplied as the outputs of voice frequency, frequency related amplitude, and voice fundamental frequency.

3.2.3. Feature Extraction

Implementation of feature extraction could be regarded as the mathematical calculation of voice frequency and voice fundamental frequency as introduced in Section 2.2.1 and 2.2.2 respectively. The step was done before classification and after pre-processing. There were totally fifteen features calculated, they are: Mean frequency (f_{mean}), Standard Deviation (f_{sd}), Median frequency (f_{median}), First Quartile (Q_{25}), Third Quartile (Q_{75}), Interquartile range (Q_{IQR}), Skewness ($f_{skewness}$), Kurtosis ($f_{kurtosis}$), Spectral Entropy (f_{spent}), Spectral Flatness (f_{sfm}), Mode frequency (f_{mode}), Frequency centroid ($f_{centroid}$), Mean fundamental frequency (F_{mean}), Minimum fundamental frequency (F_{min}), Maximum fundamental frequency (F_{max}).

Figure 9 presented extraction feature function step in Matlab and indicated these 15 features which were calculated according to the equations in Chapter 2.

FeatureExtraction.m*

```
function output = FeatureExtraction(signal,samplingrate)
%%%%Obtain voice frequency and voice fundamentl frequency from pre-preocessing file
[Frequency_amp, Frequency,frequency,FunFrequency] = PreProcessing(signal,samplingrate);

%%%Mean Frequency
meanfrequency = meanfreq(signal,samplingrate,[20 280])/1000;
%%%Standard Deviation
sd = std(Frequency);
%%%Median Frequency
median = medfreq(signal,samplingrate,[20 280])/1000;
%%%First Quartile
Q25 = quantile( Frequency,0.25);
%%%Third Quartile
Q75 = quantile( Frequency,0.75);
%%%Interquartile Range
IQR = iqr( Frequency);
%%%Skewness
skew = skewness(frequency);
%%%Kurtosis
kurt = kurtosis(frequency);
%%%%Spectral Entropy
spent = -sum(Frequency_amp.*log(Frequency_amp))./log(length(Frequency));
%entropy of a frequency spectrum:S=-sum(ylogy)/log(N)
%%%%Spectral Flatness
sfm = geomean(Frequency)/mean(Frequency);
%the ratio between the geometric mean and the arithmetic means
%%%%Mode Frequency
modefreq = mode(Frequency);
%%%Frequency Centroid
centroid = sum(Frequency.*Frequency_amp);
%%%mean fundamental frequency
meanfun = mean(FunFrequency);
%%%%minimum fundamental frequency
minfun = min(FunFrequency);
%%%maximum fundamental frequency
maxfun = max(FunFrequency);

output = [meanfrequency,sd,median,Q25,Q75,IQR,...
    skew,kurt,spent,sfm,modefreq,centroid,meanfun,minfun,maxfun];
end
```

Figure 9. Feature Extraction in Matlab

3.2.4. Classification

Multiclass Naïve Bayes classifier implemented the gender classification task, which was done after feature extraction. As the procedure indicated in Figure 8, Kory Becker's voice features dataset[10], is a public processed feature dataset created for identifying the gender of a voice which will be briefly introduced in the following section. All the features with a corresponding class label of a training set were trained into the training model by Matlab function command '*fitcnb*'. After the fifteen extracted voice features were input into the pattern matching model, they would be then classified by Naïve Bayes classifier according to the training model, which was named as *BayesModel* in Matlab.

Finally, a final decision whether this speech was from a male or a female would be made by Naïve Bayes classifier. In Matlab, the prediction work was accomplished by the function command '*BayesModel.predict'*.

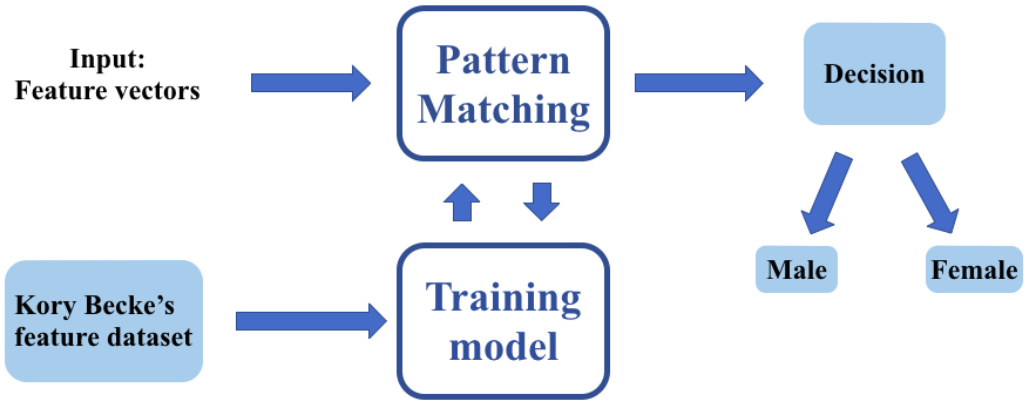

Figure 10. Demonstration of classification model

3.3. Training set

To analyze gender by voice and speech, a training database was required into training model. Kory Becke's database was built using 3168 samples of male and female voices, which were all labelled by their gender information. These voice samples were collected from the institution: The Harvard-Haskins Database of Regularly-Timed Speech [11], Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University [12], VoxForge Speech Corpus [13], Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University[14].

[10] A voice feature database made by Kory Becker.
[11] A public voice corpus resources.
[12] A public voice corpus resources.
[13] A public voice corpus resources.
[14] A public voice corpus resources.

Kory Becke's feature database contains 3168 rows and 21 columns ('meanfreq', 'sd', 'median', 'Q25', 'Q75', 'IQR', 'skew', 'kurt', 'sp.ent', 'sfm', 'mode', 'centroid', 'meanfun', 'minfun', 'maxfun', 'meandom', 'mindom', 'maxdom', 'dfrange', 'modindx', 'label'). The first 20 columns record each individual feature and the last one is label column for the gender classification. Figure 11 demonstrated each feature value in the Kory Beckes's feature database. Among the 1584 samples, the red color refers to feature values labeled as female and the blue one as male.

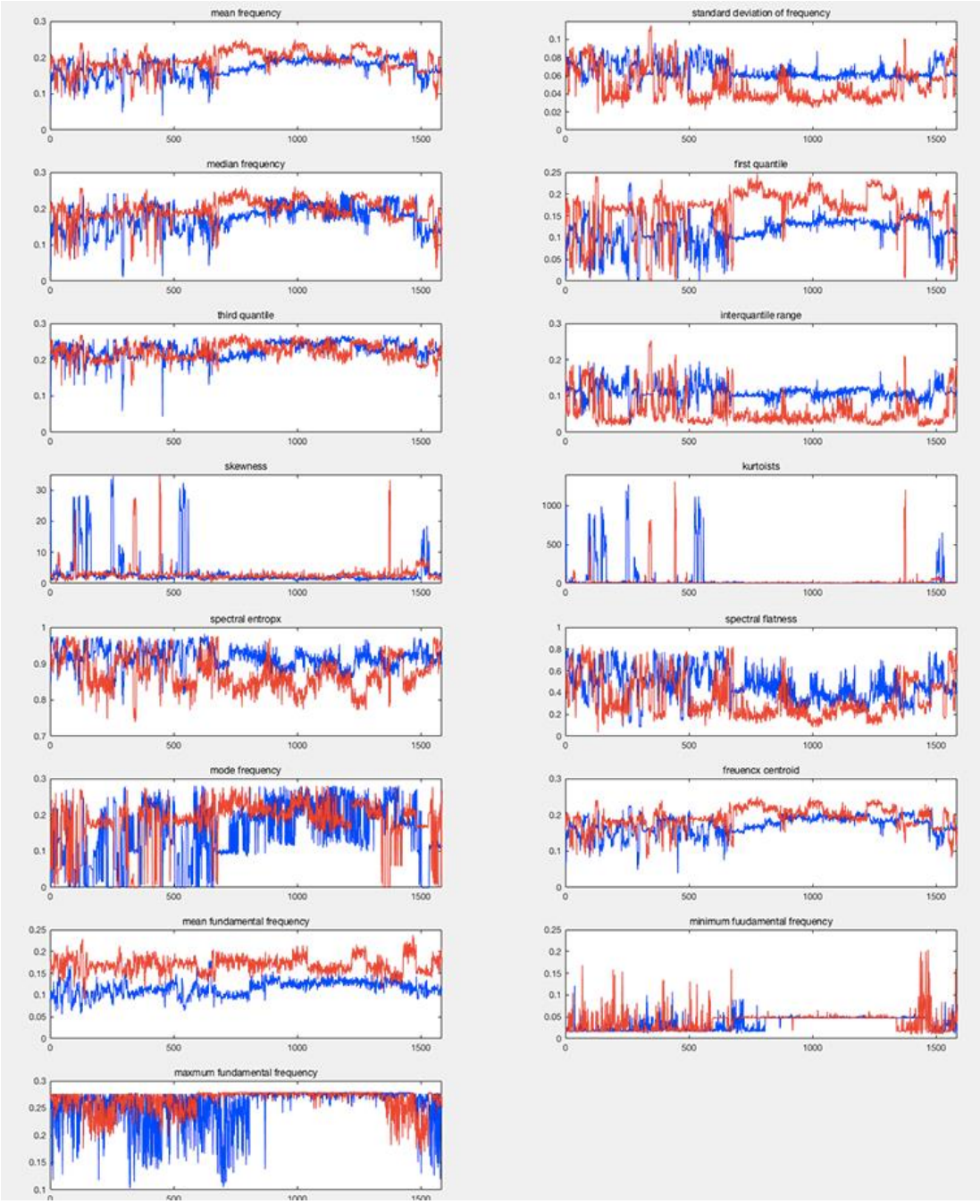

Figure 11.Training set

In this speaker recognition system, only the first fifteen feature of Kory Becke's database was extracted. The label 'Male' was replaced by number '1' and label 'Female' by number '2' so that the information could be read by Matlab. In this training set, the first 1584 rows with 15 feature columns belong to male with label '1', and the left rows from 1585 to 3168 with 15 feature columns belong to the female with label '2'.

3.4. Graphical User Interface

A graphical user interface (GUI) of speaker recognition system was designed as shown in Figure 12. All the procedure such as voice recording, pre-processing, feature extraction and classification was run in the backstage.

Voice recording can be realized through real-time voice recording function by clicking 'Start Record' Button and 'Stop Record' button, or choose a local '.wav' format file by clicking 'Choose a local record' button. The recorded speech can be presented by clicking the 'Play record' button.

'Show result' allowed a backstage operating procedure, such as pre-processing, feature extraction, and classification. Therefore, a time-domain voice graphic and fifteen feature values would be shown in GUI. Finally, a gender decision could be made and presented as well.

Additionally, the button 'Clear' can remove the recently saved voice and all the data shown in GUI, while the 'Exit' button could close this GUI.

Figure 12. Demonstration of GUI

4. EXPERIMENT AND RESULT

In this chapter, an experiment was designed to evaluate practical implementation work of the speaker gender recognition system as presented in Chapter 3. This experiment aimed at using the speaker gender recognition system to test the gender by voice input from 40 persons. To begin with, the background of this experimental voice recording and the collected experimental speech database was introduced. After this, the result of this experiment and analysis was described.

4.1. Experimental speech database

This experimental work was implemented at a quiet environment almost without noise. The voice recording equipment is Matlab in macOs version, and recorder handset is a microphone. The recording duration time was varied from person to person, therefore, sampling numbers of each voice were different. The default sampling rate is 11025 Hz with 16bits per sample and default channel number was set as 1.

There were totally 40 participants from 16 countries participated in this experiment, and 20 of them are male, and 20 are female. Additionally, the ages of all the participants range between 18 and 35, studying in University of Oulu in different faculties.

During the experiment, each participant was required to speak the following five sentences with their personal information in the blank space as follows:

My name is ____.
I am a (Female/Male).
I am ___ years old.
I come from _____ (Nationality).
I study in university of Oulu.

Therefore, the recorded voice contains information of participant's name, gender, age, nationality. In the experiment recording step, each participant was required to record ten times of the same sentences, five times in English and five times in their mother language. Since there were 40 persons involved in the experiment, 400 speech samples were totally obtained.

This experimental speech database includes 400 speech samples in 17 languages: English, Finnish, Chinese, Hindi, Sinhala, Bengali, Vietnam, Mozambique, Guinea language, Korean language, Pakistan language, Bulgarian, Mexico, French, Turkish, Russian, and Japanese. Figure 13 shows the number of different language speech samples in the experimental speech database. It can be seen that both male speech database and female speech database contain 100 speech samples respectively. The left 100 male speeches distributed in language: Finnish, Chinese, Hindi, Sinhala, Bengali, Vietnam, Mozambique, Guinea, Korean, Pakistan. The left 100 female speeches distributed in different languages: Finnish, Chinese, Bengali, Bulgarian, Mexico, French, Turkish, Russian, and Japanese.

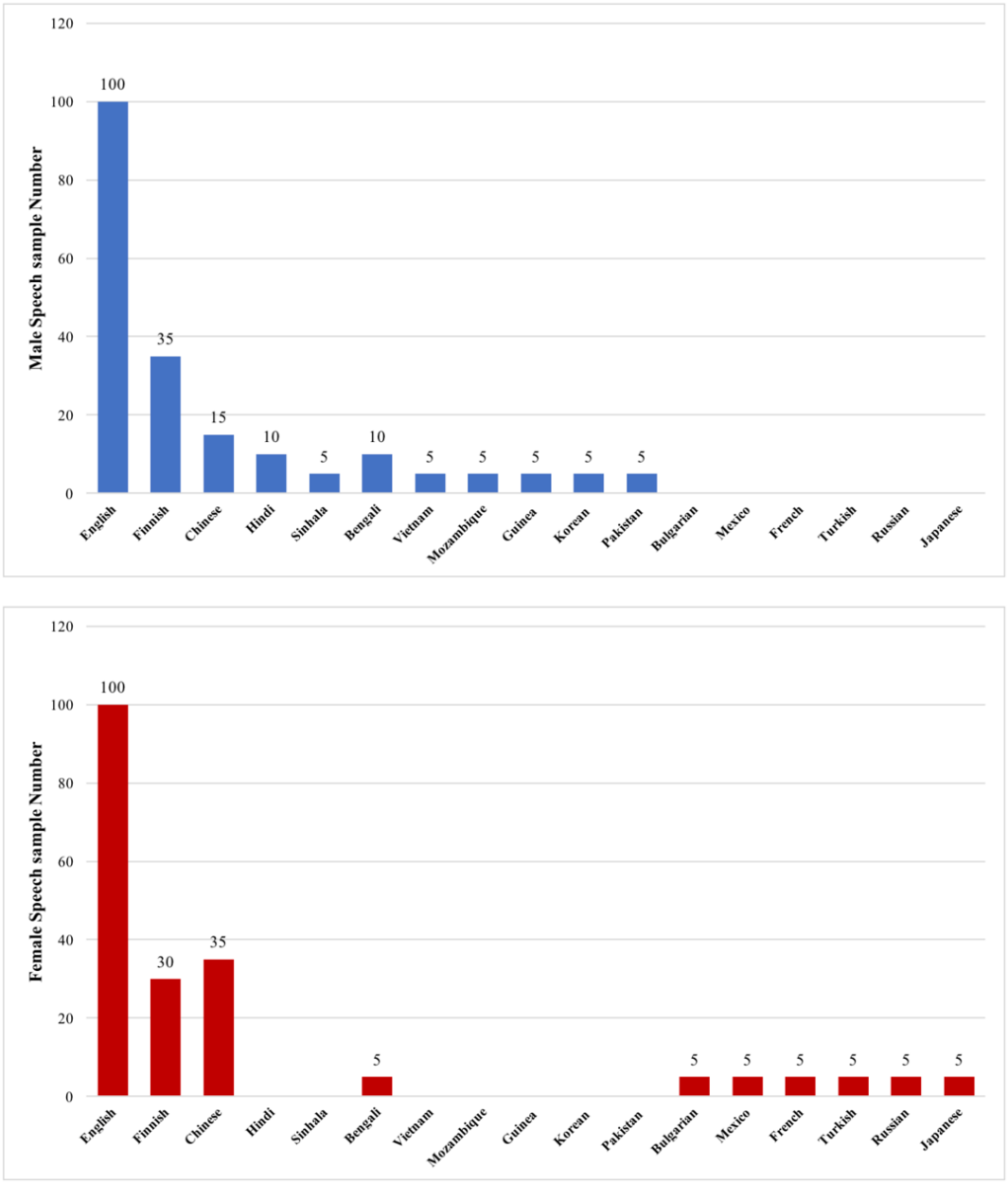

Figure 13. Experimental speech database

4.2. Result and Analysis

This section mainly analyzed the experimental result of this practical implementation work of speaker gender recognition system by using experimental speech database obtained in Section 4.1. All the results were analyzed in this section.

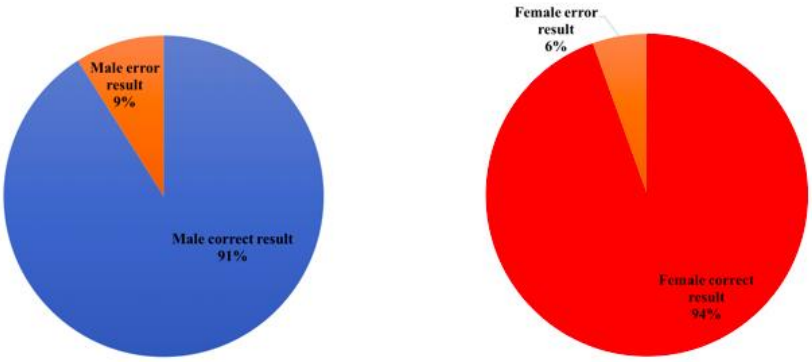

Figure 14. Speaker gender recognition experiment result

After the input of 400 speech samples into the gender recognition system, the overall accuracy of all speech samples reached 92.65%, which means that only 29 samples was not recognized correctly. Among the 29 error results, 18 of them are male voice and 11 of them are from females, as shown in Figure 14. Therefore, the accuracy of testing male voice and female voice is 91% and 94%, respectively. In other words, the accuracy of the male voice, female voice and total voice are all higher than 90%. In conclusion, the as designed speaker recognition system in this thesis is reliable.

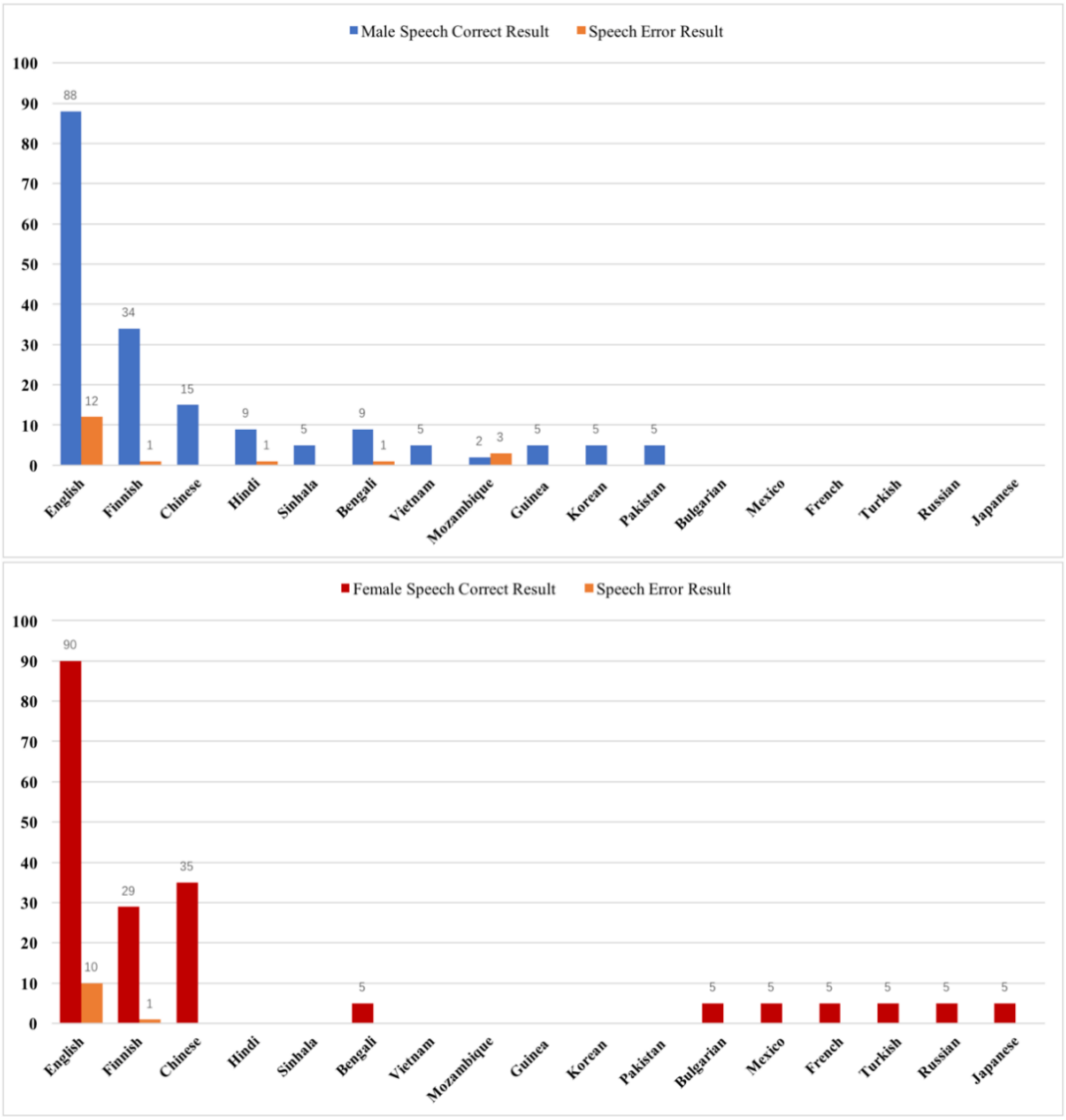

Figure 15.Experiment result in different language

The language relative accuracy is indicated by male speech correct number, female speech correct number and speech error number respectively in the different language as shown in Figure 15. It clearly shows that most of the error recognition happens with English language (12 and 10 speech error results in English male speech samples and English female speech samples respectively). There is also 3 speech error results with Mozambique language, which also takes the highest error rate (60%) language in this experiment. However, the total speech sample number of Mozambique language is 5, and the error rate cannot strictly prove that the Mozambique always obtain the lowest accuracy by using this speaker gender recognition system. Furthermore, 3 Mozambique language error results are from the same person, in other words, it might be also caused by individual condition or other interference factors.

In addition, there are also one error result from Hindi male speech, one from Finnish female speech, and one from Finnish male speech. In this experiment, accuracy is the same (90%) for English female speakers, Bengali male speakers,

Finnish male speakers and Finnish female speakers. The accuracy of Mozambique male speaker is 40%, and the accuracy of English male speaker is 88%. The accuracy of the other 12 language is 100%.

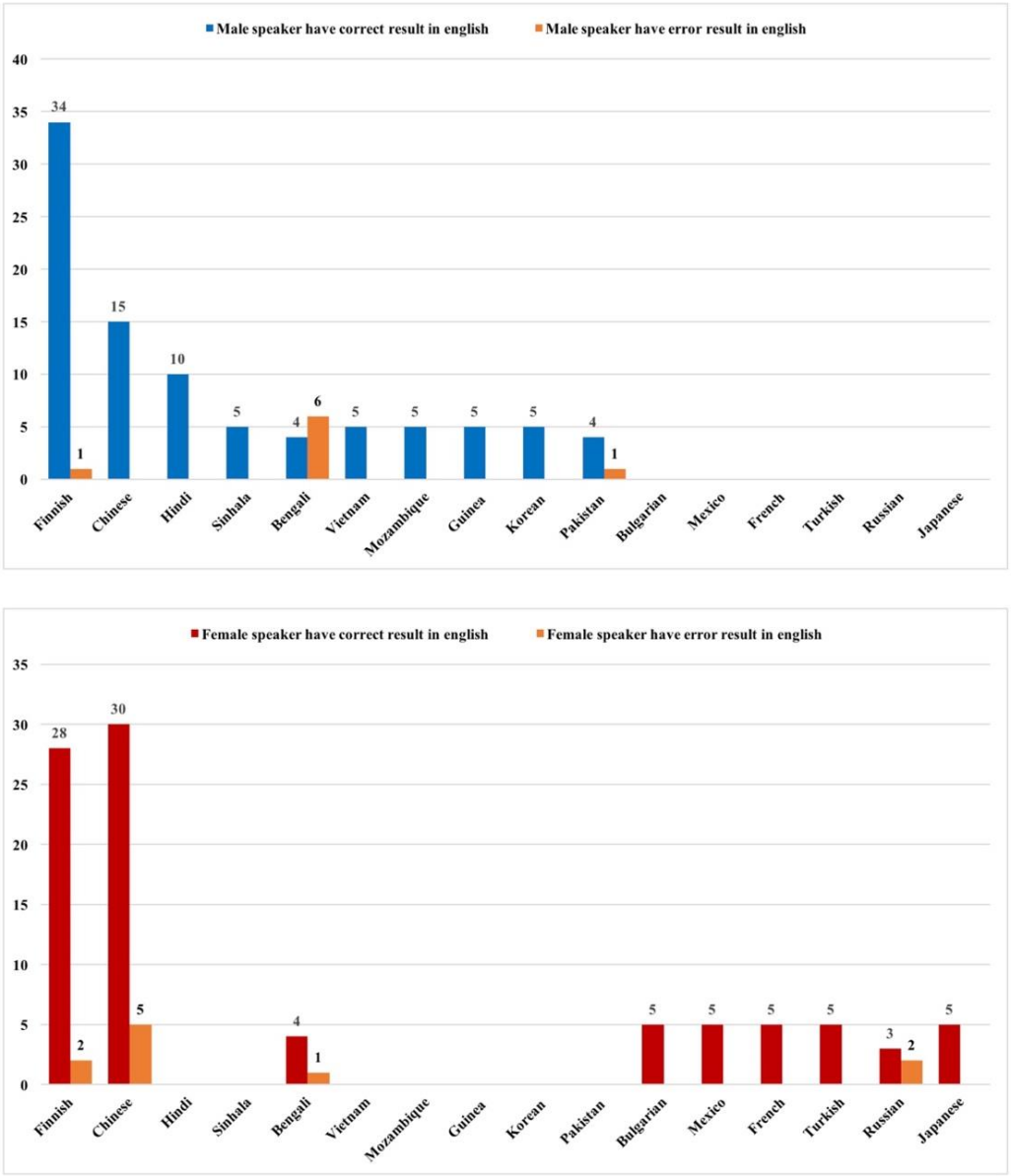

Figure 16. English speech result sorted by the nationality

By analyzing the experiment gender results in English speech effect by the nationality of speakers (or participants), Figure 16 extracted all the English speech gender results (both correct and error results) which were classified by the nationality of participants. It is clearly shown that 7 error English speech results were generated from Bengalese, including 6 from the Bengalese participant and 1 from the female participant. Therefore, we checked the information from experiment gender result database (Appendix 3), all the three Bengalese participants caused error results in this experiment.

Additionally, 5 English gender speech errors come from Chinese female speakers, and there are also 1 and 2 English gender speech errors come from Finnish male and female speakers, respectively. The left 3 English speech gender result errors are originated from a Pakistan male participant (1 error result occurs) and a Russian female participant (2 error result occurs). This irregular distribution of gender error results in English speech by speakers from different countries in this experiment implies that: the accuracy of speaker gender recognition system designed in this thesis has no affirmative correlation with the nationality background of the same spoken language.

Overall, this speaker recognition system identifies the gender of voice without the influence of the spoken languages or regional dialects. In this experiment, the speaker recognition system obtained a reliable and considerable accuracy of 92.75%.

5. DISSCUSION AND FEATURE WORK

The gender recognition by voice can be realized with multiple methods, and many speaker gender recognition systems have been designed recently. Abdulla, W. H., Kasabov, N. K., and Zealand, D. N. used the average pitch frequency of the speakers as a discriminating factor to identify the gender. They tested with TIMIT continuous speech corpus and KEL isolates words speech corpus, which showed 100% gender discrimination accuracy (no error recorded) [7]. Ali, M. S., Islam, M. S., and Hossain, M. A. in 2012 designed a system by using voice features like power spectrum density, where the frequency at maximum power carries information of speakers. They tested with 10 speakers (5 male people and 5 female people), the recognition accuracy was 80% [8]. Sheikh, H. designed a system in 2013 and investigated the different classification techniques such as Adaboost and Gaussian Mixture Models and different types of methods such as Fusion method, acoustic methods and pitch methods. They obtained an average accuracy of 80% on 10 speakers [9]. Erokyar, H. in 2014 designed an age and gender recognition system for speech application based on voice fundamental frequency and Shifted Delta Cepstral (SDC) as the voice features. They obtained an average accuracy of 64.2% with 108 experiment speakers. Ting, H., Yingchun, Y., and Zhaohui, W. (2006) combined MFCC and voice fundamental frequency in order to enhance the performance of the gender recognition. They used GMM classification and SRMC database and obtained an error rate of 3.3% [31, 32]. Harb, H., & Chen, L. (2005) designed a gender identification system which can reach an accuracy around 93% by means of a set of Neural Networks with acoustic and Pitch related features [33].

This chapter mainly discussed the performance of the designed speaker gender recognition system. To begin with, the evaluation of speaker gender recognition systems in this thesis was compared with other designed voice gender recognition systems. After that, a detailed comparison of distinct classification applied in speaker gender recognition systems were performed. Finally, the prospect in this research field was preliminarily discussed.

5.1. Evaluation of designed system

The accuracy of recognition can mainly represent the performance of a recognition system. Table 1 listed the comparison of seven voice gender recognition systems (6 designed by others and 1 designed in this thesis) with the experimental accuracy results and experimental participant numbers. It is clearly shown that most voice gender recognition systems can reach an accuracy of 80%. The designed voice gender recognition system of Harb, H., & Chen, L. (2005) and Ting, H., Yingchun, Y. and Zhaohui, W. (2006) obtained relatively high recognition accuracy result over 90%. Abdulla, W. H., Kasabov, N. K., & Zealand, D. N. (2001) even obtain errorless accuracy with their recognition system.

However, in the report by Abdulla, W. H., Kasabov, N. K., & Zealand, D. N., there was no real participants involved in the test although their recognition system achieved totally accurate result. Therefore, it is hard to prove that their system can always reach 100% accuracy. Harb, H., & Chen, L. reached 93% recognition accuracy while the involved 4 French radio station and 1 English radio station were not persuadable in case of real speakers. In our work, the recognition accuracy of

designed speaker recognition system is 92.75% with 40 speaker participants (20 male people and 20 female people). Even though the accuracy obtained from this thesis is slightly lower than those obtained from the work done by Ting, H., Yingchun, Y., & Zhaohui, W (96.7%), there are much more participants in this thesis. Compared with the work done by Ali, M. S., Islam, M. S., & Hossain, M. A. (2012) and Sheikh, H. (2013), the speaker gender recognition system performed better in both recognition accuracy and scope of the tests. The recognition accuracy of Erokyar, H. is only 64.2%, but it combined both age recognition and gender recognition. The age recognition might be the work which will attract more attention in the future.

Thus, even our recognition accuracy and scope of experiment is not the best one comparing with the other six voice gender recognition systems, it is still shows a promising performance with consideration of both recognition accuracy and the quantity of experiment participants.

Table 1. Comparison of different gender recognition system result.

Authors	**Year**	**Accuracy**	**Number of Participants**
Abdulla, W. H., Kasabov, N. K., & Zealand, D. N. [7]	2001	100%	TIMIT continuous speech corpus and KEL isolate words speech corpus
Harb, H., & Chen, L. [33]	2005	93%	4 French radio station, 1 English radio station
Ting, H., Yingchun, Y., & Zhaohui, W. [31]	2006	96.7%	10 male and 10 female
Ali, M. S., Islam, M. S., & Hossain, M. A. [8]	2012	80%	10
Sheikh, H. [9]	2013	Almost 80%	10
Erokyar, H. [10]	2014	Average 64.2% (With Gender)	108
My work	2017	92.75%	40

According to the trend of the aforementioned researches, machine learning method is a tendency in the application of speaker gender recognition system. As shown in Table 2, the built of machine learning gender recognition system requires voice features, classifications, training database. It illustrates that voice fundamental frequency related features are widely used as extracted voice features. In addition, MFCC also provides a good performance for voice gender recognition, thus MFCC might be incorporated in our future research (Section 5.3). Since there is no obvious conclusion to choose the best classification for speaker gender recognition, the influence of different classification methods on the recognition accuracy should be considered, and a related comparison work has been done in Section 5.2.

Combing with previous research and the experiment with 17 different language in this thesis, it can fully be explained that language is not a limitation factor for the gender recognition accuracy according to a speech.

Table 2. Machine learning method used in voice gender recognition system

	Ting, H., Yingchun, Y., & Zhaohui, W.	**Harb, H., & Chen, L.**	**Work done in this thesis**
Voice Features	MFCC, Voice fundamental frequency	Acoustic, Voice fundamental frequency	Voice frequency related Voice fundamental frequency related
Classification	GMM	Neural Network	Naïve Bayes
Training Database	SRMC [32]	Switchboard	Kory Becke's
Tested Speech Language	English, Chinese	English, French	English, Finnish, Chinese, Hindi, Sinhala, Bengali, Vietnam, Mozambique, Guinea, Korean, Pakistan, Bulgarian, Mexico, French, Turkish, Russian, Japanese

5.2. Comparison of Classification

In our designed speaker gender recognition system, the Naïve Bayes classifier was used as the machine learning classification method. In this section, we apply another classification method to compare the recognition accuracy variation with different classification method.

In the classification step, we replace other four classification methods by Naïve Bayes classification without changing all the other steps in Matlab, however, the experimental speech database was kept unchanged as shown in Section 4.1. The other four classifications are Support Vector Machines classification, tree classification, A nearest-neighbor (KNN) classification and error-correcting output codes (ECOC). Finally, we get recognition accuracy results shown in Figure 17.

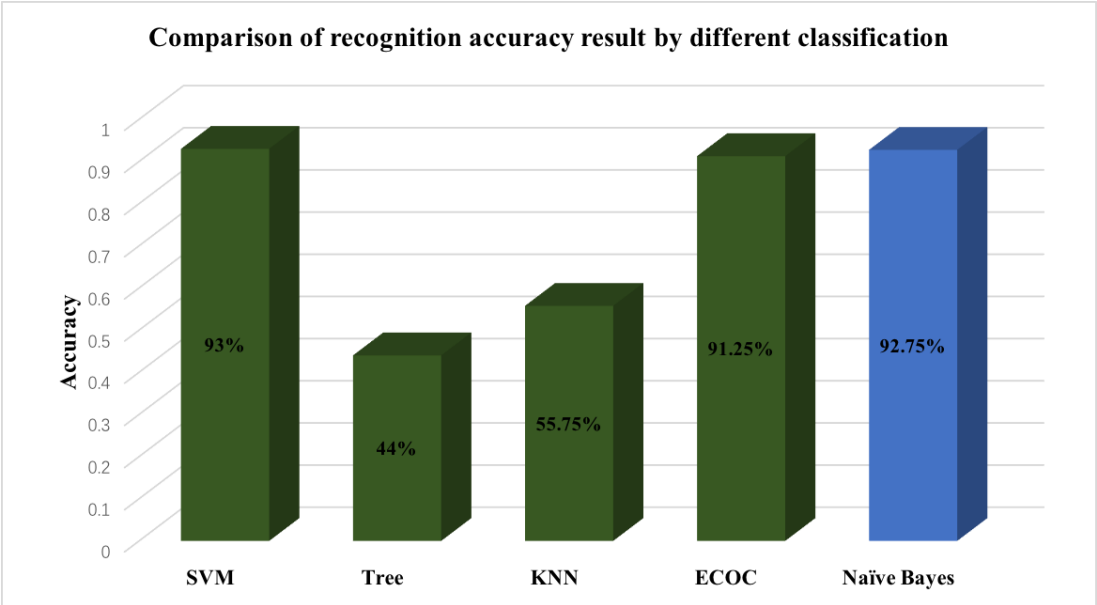

Figure 17.Comparison of recognition accuracy result by different classification

From Figure 17, we can see that SVM classifier obtained slightly higher recognition accuracy (93%) than that by Naïve Bayes classifier (92.75%). ECOC classifier provides a close accuracy (91.25%) to Naïve Bayes classifier. Tree classification model and KNN classification model caused a very worse recognition accuracy are 44% and 55.75% respectively.

In general, the Naïve Bayes classifier provide good performance compared with the other four classifiers. Even though SVM classifier brought a slightly improved accuracy of 0.25% than Naïve Bayes classifier, there is no persuaded proof to claim the SVM as a better classifier than Naïve Bayes classifier.

In the future, we might design another experiment to compare the influence of classification methods.

5.3. Future work

For the future research, I would like to improve the accuracy of gender recognition accuracy. The performance of different classification would be compared to obtain a conclusion whether different classification effected the accuracy result of a machine learning voice gender recognition. And the classification method with the best performance will be used in the speaker recognition system.

In addition, we would like to create a perfect speaker recognition system with multiple functionalities. The research of age recognition function by using the same speech would be our core work in the next step. After this, the research work will be focused on emotion recognition and region recognition. We aimed at designing a speaker recognition system which might be applied in IoT, such as user authentication by voice, where the voice instead of keyboard input will be the input commands. The final goal of this thesis is design a system applied as a part of voice authentication into the Naked Approach project.

I have to acknowledge that there are still several limitation factors in this thesis work. Although a 400 speech samples database is collected in 17 languages from 40

participants including both male speakers and female speakers, the number of participants is not enough to extensively prove the recognition results influenced by languages and regional dialects. Besides, all the participants in this experiment are in the young ages ranging between 18 and 35. Therefore, we need an experimental speech database with more participants, extended age range and expanded nationalities.

6. SUMMARY

The objective of this thesis is to build a speaker gender recognition system. This thesis introduced all the necessary technical backgrounds and the steps required to establish such a system. The technical backgrounds in this thesis include voice pre-processing, voice feature extraction and classification. This system is able to identify the gender of a person under a quite environment by recording and processing the speech of this tested person.

In this thesis, the practical implementation steps of the built design includes voice recording, voice pre-processing, voice feature extraction and classification. This implementation work is finished with Matlab and the final speaker gender recognition system can be demonstrated as a graphic user interface (Figure 12). This graphic user interface presents a real-time speaker gender recognition system which could identify the speech from a microphone or a recorded speech file in .wav format.

This performance of the designed gender identification system was evaluated by an experiment of 40 people from 16 countries. The environment of this experiment is a quite small room without noise. Finally, all the recorded speeches constituted an experimental speech database containing 400 speech samples from the male and the female voice in 17 languages.

According to the test results, the total average recognition accuracy is 92.75%. Furthermore, we also analyzed the influence of different languages and regional accents on this recognition system. The tested results indicates that there is no obvious influence from the spoken language or accent to this built system.

Moreover, we compared the recognition accuracy obtained by this system with several other speaker gender recognition systems. From the comparison results, we can conclude that the performance of the built speaker recognition system in this thesis is good enough. In addition, we also discussed the performance of different classification to this system. We can see that the classification we chose in the built systems provides a higher accuracy than the others, however, SVM classification reported a slightly improved accuracy of 0.25% but without actual participants. To sum up, the speaker recognition system designed in this thesis has a reliable and considerably good performance for identifying the gender of a speaker without the limitation of their nationalities and languages.

In the future work, we plan to improve the gender recognition accuracy of this system. The best classification way would be selected and applied into the improved system. A larger scale experimental database might be also built with more participants from more countries in different age groups. The future research will focus on more functionalities such as age recognition, nationality recognition, and emotion recognition.

7. REFERENCES

[1] Burnett, M., & Kulesza, T. (2015) End-User Development in Internet of Things: We the People. In *International Reports on Socio-Informatics (IRSI)*, Vol.12, Iss.2, pp. 81-86.

[2] Childers, D. G., Wu, K., Bae, K. S., & Hicks, D. M. (1988, April). Automatic recognition of gender by voice. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on* (pp. 603-606). IEEE.

[3] Sigmund, M. (2008). Gender distinction using short segments of speech signal. *International Journal of Computer Science and Network Security*, *8*(10), 159-162.

[4] Tolba, H. (2011). A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach. *Alexandria Engineering Journal*, *50*(1), 43-47.

[5] Chisaki, Y., Nakashima, H., Shiroshita, S., Usagawa, T., & Ebata, M. (2003). A pitch detection method based on continuous wavelet transform for harmonic signal. *Acoustical science and technology*, *24*(1), 7-16.

[6] Sigmund, M. (2008). Gender distinction using short segments of speech signal. *International Journal of Computer Science and Network Security*, *8*(10), 159-162.

[7] Abdulla, W. H., Kasabov, N. K., & Zealand, D. N. (2001). Improving speech recognition performance through gender separation. *changes*, *9*, 10.

[8] Ali, M. S., Islam, M. S., & Hossain, M. A. (2012). Gender recognition system using speech signal. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, *2*(1), 1-9.

[9] Sheikh, H. (2013). Who is Speaking? Male or Female.

[10] Bissell, C. C. (1990). Nyquist Rate Sampling. *International Journal of Electrical Engineering Education*, *27*(1), 77-79.

[11] Battu, D. (2014). New Telecom Networks: Enterprises and Security. *John Wiley & Sons*, pp.159-194.

[12] Grimaldi, M., & Cummins, F. (2008). Speaker identification using instantaneous frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(6), 1097-1111.

[13] Manandhar, S., Austin, J., Desai, U., Oyanagi, Y., & Talukder, A. (Eds.). (2005). Applied Computing: Second Asian Applied Computing Conference, AACC 2004, Kathmandu, Nepal, October 29-31, 2004. Proceedings (Vol. 3285). Springer.

[14] Sahoo, T. R., & Patra, S. (2014). Silence removal and endpoint detection of speech signal for text independent speaker identification. *International Journal of Image, Graphics and Signal Processing*, *6*(6), 27.

[15] Farrús, M. (2007). Jitter and shimmer measurements for speaker recognition. In 8th Annual Conference of the International Speech Communication Association; 2007 Aug. 27-31; Antwerp (Belgium). [place unknown]: ISCA; 2007. p. 778-81.. International Speech Communication Association (ISCA).

[16] Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice*, *16*(4), 480-487.

[17] Mathcentre. (2003). Variance and standard deviation (grouped data)**.**

[18] Kris, M & Timothy, P. (2010). *Five-Number Summary and Box-and-Whisker Plots*. SoftChalk LessonBuilder.

[19] Blanco, S., Garay, A., & Coulombie, D. (2013). Comparison of frequency bands using spectral entropy for epileptic seizure prediction. *ISRN neurology*, *2013*.

[20] Gerhard, D. (2003). *Pitch extraction and fundamental frequency: History and current techniques* (pp. 0-22). Regina: Department of Computer Science, University of Regina.

[21] De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917-1930.

[22] McLeod, P., & Wyvill, G. (2005, September). A Smarter Way to Find pitch. In *ICMC*.

[23] Middleton, G. (2003). Pitch detection algorithms.

[24] Noll, A. M. (1967). Cepstrum pitch determination. *The journal of the acoustical society of America*, *41*(2), 293-309.

[25] Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, *123*(6), 4559-4571.

[26] Titze, I. R., & Martin, D. W. (1998). Principles of voice production. *The*

[27] James, L. (2009-2012). Mel Frequency Cepstral Coefficient tutorial.

[28] Murty, M. N., & Devi, V. S. (2011). *Pattern recognition: An algorithmic approach*. Springer Science & Business Media.

[29] Huang, W., & MacFarlane, D. L. (2016). Fast Fourier Transform and MATLAB Implementation. *The University of Texas at Dallas. Dr. Duncan L. MacFarlane. Web*, *24*.

[30] Abdulla, W. H., Kasabov, N. K., & Zealand, D. N. (2001). Improving speech recognition performance through gender separation. *changes*, *9*, 10.

[31] Ting, H., Yingchun, Y., & Zhaohui, W. (2006). Combining MFCC and pitch to enhance the performance of the gender recognition. In *Signal Processing, 2006 8th International Conference on* (Vol. 1). IEEE.

[32] Sang, L., Wu, Z., & Yang, Y. (2003, October). Speaker recognition system in multi-channel environment. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on* (Vol. 4, pp. 3116-3121). IEEE.

[33] Harb, H., & Chen, L. (2005). Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, *24*(2), 179-198.

8. APPENDICES

Appendix 1 Research Consent Form
Appendix 2 Participants' information
Appendix 3 Experimental Results

Appendix 1 Research Consent Form

Research Briefing

This test records the voice of participant speaking some sentences for 10 to 15 times. The recorded voice will be published as public information and used for the evaluation of Hong Zimeng's master thesis, but the private information of participant will be confidential.

The duration of the experiment is approximately 5 to 10 minutes. The participants will be compensated with small gifts.

Please speak following sentence for 5 times:
My name is _____.
I am a (Female/Male).
I am _____ years old.
I come from _____. (Nationality)
I study in university of Oulu.

Consent Form

- I agree that Project Team can audio record my test session for evaluation analysis.
- I understand that any information collected during this evaluation will be treated as confidential.
- I understand that my information will be destroyed after the completion of the project.

I have read all the information above and I [ ] agree / [ ] disagree
Date: _____ Signature: _____
Email: _____

Appendix 2 Participants' information

Participant	Gender	Age	Nationality
1	Male	25	Finland
2	Male	24	Finland
3	Male	24	China
4	Male	23	China
5	Male	35	Finland
6	Male	24	India
7	Male	21	Sri Lanka
8	Male	28	India
9	Male	24	Bangladesh
10	Male	19	Finland
11	Male	20	Finland
12	Male	21	Mozambique
13	Male	25	Finland
14	Male	25	Guinea
15	Male	20	Vietnam
16	Male	24	Korea
17	Male	25	Finland
18	Male	25	Bangladesh
19	Male	26	Pakistan
20	Male	25	China
21	Female	24	Finland
22	Female	22	Finland
23	Female	21	China
24	Female	21	China
25	Female	19	China
26	Female	25	Finland
27	Female	23	Bulgaria
28	Female	29	Bangladesh
29	Female	23	China
30	Female	21	Mexico
31	Female	20	China
32	Female	21	France
33	Female	22	China
34	Female	21	Turkey
35	Female	21	China
36	Female	21	Finland
37	Female	23	Russia
38	Female	24	Finland
39	Female	29	Japan
40	Female	31	Finland

Appendix 3 Experimental Results

Sample Number	*Gender*	*Language*	*Result*	*Sample Number*	*Gender*	*Language*	*Result*
1	Male	English	correct	**201**	Female	English	correct
2	Male	English	correct	**202**	Female	English	correct
3	Male	English	correct	**203**	Female	English	correct
4	Male	English	correct	**204**	Female	English	correct
5	Male	English	correct	**205**	Female	English	correct
6	Male	Finnish	correct	**206**	Female	Finnish	correct
7	Male	Finnish	correct	**207**	Female	Finnish	correct
8	Male	Finnish	correct	**208**	Female	Finnish	correct
9	Male	Finnish	correct	**209**	Female	Finnish	correct
10	Male	Finnish	correct	**210**	Female	Finnish	correct
11	Male	English	correct	**211**	Female	English	correct
12	Male	English	correct	**212**	Female	English	correct
13	Male	English	correct	**213**	Female	English	correct
14	Male	English	correct	**214**	Female	English	correct
15	Male	English	correct	**215**	Female	English	correct
16	Male	Finnish	correct	**216**	Female	Finnish	correct
17	Male	Finnish	correct	**217**	Female	Finnish	correct
18	Male	Finnish	correct	**218**	Female	Finnish	correct
19	Male	Finnish	correct	**219**	Female	Finnish	correct
20	Male	Finnish	correct	**220**	Female	Finnish	correct
21	Male	English	correct	**221**	Female	English	correct
22	Male	English	correct	**222**	Female	English	correct
23	Male	English	correct	**223**	Female	English	correct
24	Male	English	correct	**224**	Female	English	correct
25	Male	English	correct	**225**	Female	English	correct
26	Male	Chinese	correct	**226**	Female	Chinese	correct
27	Male	Chinese	correct	**227**	Female	Chinese	correct
28	Male	Chinese	correct	**228**	Female	Chinese	correct
29	Male	Chinese	correct	**229**	Female	Chinese	correct
30	Male	Chinese	correct	**230**	Female	Chinese	correct
31	Male	English	correct	**231**	Female	English	correct
32	Male	English	correct	**232**	Female	English	correct
33	Male	English	correct	**233**	Female	English	correct
34	Male	English	correct	**234**	Female	English	correct
35	Male	English	correct	**235**	Female	English	correct
36	Male	Chinese	correct	**236**	Female	Chinese	correct
37	Male	Chinese	correct	**237**	Female	Chinese	correct
38	Male	Chinese	correct	**238**	Female	Chinese	correct
39	Male	Chinese	correct	**239**	Female	Chinese	correct
40	Male	Chinese	correct	**240**	Female	Chinese	correct
41	Male	English	correct	**241**	Female	English	error
42	Male	English	correct	**242**	Female	English	correct

43	Male	English	correct	**243**	Female	English	correct
44	Male	English	correct	**244**	Female	English	correct
45	Male	English	correct	**245**	Female	English	correct
46	Male	Finnish	correct	**246**	Female	Chinese	correct
47	Male	Finnish	correct	**247**	Female	Chinese	correct
48	Male	Finnish	correct	**248**	Female	Chinese	correct
49	Male	Finnish	correct	**249**	Female	Chinese	correct
50	Male	Finnish	correct	**250**	Female	Chinese	correct
51	Male	English	correct	**251**	Female	English	correct
52	Male	English	correct	**252**	Female	English	correct
53	Male	English	correct	**253**	Female	English	correct
54	Male	English	correct	**254**	Female	English	correct
55	Male	English	correct	**255**	Female	English	correct
56	Male	Hindi	correct	**256**	Female	Finnish	correct
57	Male	Hindi	correct	**257**	Female	Finnish	correct
58	Male	Hindi	correct	**258**	Female	Finnish	correct
59	Male	Hindi	correct	**259**	Female	Finnish	correct
60	Male	Hindi	correct	**260**	Female	Finnish	correct
61	Male	English	correct	**261**	Female	English	correct
62	Male	English	correct	**262**	Female	English	correct
63	Male	English	correct	**263**	Female	English	correct
64	Male	English	correct	**264**	Female	English	correct
65	Male	English	correct	**265**	Female	English	correct
66	Male	Sinhala	correct	**266**	Female	Bulgarian	correct
67	Male	Sinhala	correct	**267**	Female	Bulgarian	correct
68	Male	Sinhala	correct	**268**	Female	Bulgarian	correct
69	Male	Sinhala	correct	**269**	Female	Bulgarian	correct
70	Male	Sinhala	correct	**270**	Female	Bulgarian	correct
71	Male	English	correct	**271**	Female	English	correct
72	Male	English	correct	**272**	Female	English	error
73	Male	English	correct	**273**	Female	English	correct
74	Male	English	correct	**274**	Female	English	correct
75	Male	English	correct	**275**	Female	English	correct
76	Male	Hindi	correct	**276**	Female	Bengali	correct
77	Male	Hindi	error	**277**	Female	Bengali	correct
78	Male	Hindi	correct	**278**	Female	Bengali	correct
79	Male	Hindi	correct	**279**	Female	Bengali	correct
80	Male	Hindi	correct	**280**	Female	Bengali	correct
81	Male	English	correct	**281**	Female	English	correct
82	Male	English	error	**282**	Female	English	correct
83	Male	English	error	**283**	Female	English	correct
84	Male	English	error	**284**	Female	English	correct
85	Male	English	correct	**285**	Female	English	correct
86	Male	Bengali	correct	**286**	Female	Chinese	correct
87	Male	Bengali	correct	**287**	Female	Chinese	correct
88	Male	Bengali	correct	**288**	Female	Chinese	correct
89	Male	Bengali	correct	**289**	Female	Chinese	correct

90	Male	Bengali	correct	**290**	Female	Chinese	correct
91	Male	English	correct	**291**	Female	English	correct
92	Male	English	correct	**292**	Female	English	correct
93	Male	English	correct	**293**	Female	English	correct
94	Male	English	correct	**294**	Female	English	correct
95	Male	English	correct	**295**	Female	English	correct
96	Male	Finnish	correct	**296**	Female	Mexico	correct
97	Male	Finnish	error	**297**	Female	Mexico	correct
98	Male	Finnish	correct	**298**	Female	Mexico	correct
99	Male	Finnish	correct	**299**	Female	Mexico	correct
100	Male	Finnish	correct	**300**	Female	Mexico	correct
101	Male	English	correct	**301**	Female	English	correct
102	Male	English	correct	**302**	Female	English	correct
103	Male	English	error	**303**	Female	English	correct
104	Male	English	correct	**304**	Female	English	error
105	Male	English	correct	**305**	Female	English	error
106	Male	Finnish	correct	**306**	Female	Chinese	correct
107	Male	Finnish	correct	**307**	Female	Chinese	correct
108	Male	Finnish	correct	**308**	Female	Chinese	correct
109	Male	Finnish	correct	**309**	Female	Chinese	correct
110	Male	Finnish	correct	**310**	Female	Chinese	correct
111	Male	English	error	**311**	Female	English	correct
112	Male	English	error	**312**	Female	English	correct
113	Male	English	error	**313**	Female	English	correct
114	Male	English	correct	**314**	Female	English	correct
115	Male	English	correct	**315**	Female	English	correct
116	Male	Chinese	correct	**316**	Female	French	correct
117	Male	Chinese	error	**317**	Female	French	correct
118	Male	Chinese	error	**318**	Female	French	correct
119	Male	Chinese	error	**319**	Female	French	correct
120	Male	Chinese	correct	**320**	Female	French	correct
121	Male	English	correct	**321**	Female	English	correct
122	Male	English	correct	**322**	Female	English	correct
123	Male	English	correct	**323**	Female	English	error
124	Male	English	correct	**324**	Female	English	error
125	Male	English	correct	**325**	Female	English	correct
126	Male	Finnish	correct	**326**	Female	Chinese	correct
127	Male	Finnish	correct	**327**	Female	Chinese	correct
128	Male	Finnish	correct	**328**	Female	Chinese	correct
129	Male	Finnish	correct	**329**	Female	Chinese	correct
130	Male	Finnish	correct	**330**	Female	Chinese	correct
131	Male	English	correct	**331**	Female	English	correct
132	Male	English	correct	**332**	Female	English	correct
133	Male	English	correct	**333**	Female	English	correct
134	Male	English	correct	**334**	Female	English	correct
135	Male	English	correct	**335**	Female	English	correct
136	Male	Guinea	correct	**336**	Female	Turkish	correct

137	Male	Guinea	correct	**337**	Female	Turkish	correct
138	Male	Guinea	correct	**338**	Female	Turkish	correct
139	Male	Guinea	correct	**339**	Female	Turkish	correct
140	Male	Guinea	correct	**340**	Female	Turkish	correct
141	Male	English	error	**341**	Female	English	correct
142	Male	English	correct	**342**	Female	English	correct
143	Male	English	correct	**343**	Female	English	correct
144	Male	English	correct	**344**	Female	English	correct
145	Male	English	correct	**345**	Female	English	correct
146	Male	Bengali	correct	**346**	Female	Chinese	correct
147	Male	Bengali	correct	**347**	Female	Chinese	correct
148	Male	Bengali	correct	**348**	Female	Chinese	correct
149	Male	Bengali	correct	**349**	Female	Chinese	correct
150	Male	Bengali	correct	**350**	Female	Chinese	correct
151	Male	English	correct	**351**	Female	English	correct
152	Male	English	correct	**352**	Female	English	correct
153	Male	English	correct	**353**	Female	English	correct
154	Male	English	correct	**354**	Female	English	correct
155	Male	English	correct	**355**	Female	English	correct
156	Male	Korean	correct	**356**	Female	Finnish	correct
157	Male	Korean	correct	**357**	Female	Finnish	correct
158	Male	Korean	correct	**358**	Female	Finnish	correct
159	Male	Korean	correct	**359**	Female	Finnish	correct
160	Male	Korean	correct	**360**	Female	Finnish	correct
161	Male	English	correct	**361**	Female	English	correct
162	Male	English	correct	**362**	Female	English	correct
163	Male	English	correct	**363**	Female	English	error
164	Male	English	correct	**364**	Female	English	correct
165	Male	English	correct	**365**	Female	English	error
166	Male	Finnish	correct	**366**	Female	Russian	correct
167	Male	Finnish	correct	**367**	Female	Russian	correct
168	Male	Finnish	correct	**368**	Female	Russian	correct
169	Male	Finnish	correct	**369**	Female	Russian	correct
170	Male	Finnish	correct	**370**	Female	Russian	correct
171	Male	English	error	**371**	Female	English	correct
172	Male	English	error	**372**	Female	English	correct
173	Male	English	correct	**373**	Female	English	correct
174	Male	English	error	**374**	Female	English	correct
175	Male	English	correct	**375**	Female	English	correct
176	Male	Bengali	correct	**376**	Female	Finnish	correct
177	Male	Bengali	correct	**377**	Female	Finnish	correct
178	Male	Bengali	correct	**378**	Female	Finnish	correct
179	Male	Bengali	error	**379**	Female	Finnish	correct
180	Male	Bengali	correct	**380**	Female	Finnish	correct
181	Male	English	error	**381**	Female	English	correct
182	Male	English	correct	**382**	Female	English	correct
183	Male	English	correct	**383**	Female	English	correct

184	Male	English	correct	**384**	Female	English	correct
185	Male	English	correct	**385**	Female	English	correct
186	Male	Pakistan	correct	**386**	Female	Japanese	correct
187	Male	Pakistan	correct	**387**	Female	Japanese	correct
188	Male	Pakistan	correct	**388**	Female	Japanese	correct
189	Male	Pakistan	correct	**389**	Female	Japanese	correct
190	Male	Pakistan	correct	**390**	Female	Japanese	correct
191	Male	English	correct	**391**	Female	English	correct
192	Male	English	correct	**392**	Female	English	correct
193	Male	English	correct	**393**	Female	English	correct
194	Male	English	correct	**394**	Female	English	error
195	Male	English	correct	**395**	Female	English	error
196	Male	Chinese	correct	**396**	Female	Finnish	correct
197	Male	Chinese	correct	**397**	Female	Finnish	correct
198	Male	Chinese	correct	**398**	Female	Finnish	correct
199	Male	Chinese	correct	**399**	Female	Finnish	error
200	Male	Chinese	correct	**400**	Female	Finnish	correct