International Journal of Innovative Computing, Information and Control
Volume **20**, Number **1**, February **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **213–229**

SHORT TIME FOURIER TRANSFORM IN REINVIGORATING DISTINCTIVE FACTS OF INDIVIDUAL SPECTRAL CENTROID OF MEL FREQUENCY NUMERIC FOR SECURITY AUTHENTICATION

Heni Ispur Pratiwi[1], Widodo Budiharto[2]
Iman Herwidiana Kartowisastro[3] and Benfano Soewito[4]

[1]Doctorate Program of Computer Science Department
[3]BINUS Graduate Program – Doctor of Computer Science Department
[4]BINUS Graduate Program – Master of Computer Science Department
Bina Nusantara University
Jl. Kebon Jeruk Raya No. 27, Jakarta 11530, Indonesia
{ heni.pratiwi001; ihkartowisastro }@binus.ac.id; bsoewito@binus.edu

[2]Computer Science Department
School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Jakarta 11480, Indonesia
wbudiharto@binus.edu

Received April 2023; revised September 2023

Abstract. *Human throat and mouth anatomy attribute to the uniqueness of a human voice, speech patterns or in Mel Scale is called spectral centroids. In general, speeches frequencies differ due to a person's age and gender with the frequency range from 0 to 5 kHz approximately. Individual speech patterns change over time as humans grow older or with illnesses which create inconsistency issues and need to be challenged to innovate. Some accuracies in identifying speakers have been achieved to offer resolutions; however, they degrade along with other issues compromised into the system, such as noises. Research stated that a person's fundamental speech is in the form of signals which they should have been computed based on voice signals. This paper displays how spectral centroid is reinvigorated by Mel Scale and Short Time Fourier Transform combined. The spectral centroid is proposed to become specific selection for login keys which can mitigate loopholes for unauthorized users undermining the authentication process. The finding of one individual's time spans of spectral centroid of [10;3804] and [20;3204] are suggested selections for login keys. This proved that unique speech patterns remain plausible being permanent or consistent and this is significant for a security system.*

Keywords: Speech recognition system, Mel Scale, Short Time Fourier Transform

1. **Introduction.** Speech Recognition System (SRS) is one of the biometric technologies that is commonly and widely used nowadays. This voice-activated based system is used to access control systems through voice commands. This technique utilizes commands at electronic devices such as stereo systems, and audio devices [1]. Since the early 1950s, SRS has been a part of research objectives and contributions used to advanced scientific literature and has been effectively applied to human-machine interaction devices in modern technology [2].

SRS in this modern era has been one of the most popular downloaded applications such as in Google voice, reservation services, home automations system, security system,

DOI: 10.24507/ijicic.20.01.213

and meeting attendance services. SRS has been widely implemented and suggested using various models in one or more methods combined [3].

There has been a phenomenon of updated new features in SRS. In the last ten years, SRS had accomplished translating isolated-words in a real-time system. Moreover, it has the ability to identify and verify speakers among many speakers randomly in various conditions and background. Developments within SRS had also contributed into increasing the accuracy rates, since increasing the system accuracies is suggested. Furthermore, a security system should not rely on accuracy performances alone. Apparently, the accuracy changes along with individual's physical conditions such as illnesses: flu, asthma, or throat cancer that could affect voice patterns changes [4]. However, the product of human anatomy sensitivity pivots seems to have been disregarded in supporting a resolution into SRS security.

The term SRS associates with Automatic Speech Recognition (ASR) which is basically an embedded SRS system. Many experiments and researches use the terms ASR when they refer to the hardware device itself. For example, there has been a research experiment to develop a processor within the ASR to increase its robustness [5]. On the other hand, when a research refers merely to the biometric speech system itself, they use the term SRS. Apart from personal authentication as an access control system, there is another result of SRS development that is a device called Automatic Speech Verification or ASV [6].

There are issues in SRS that cause more complicated security problems, especially when the system is compromised by unauthorized users. Technically, the origin data are in the form of voice signals used as input to the system and could be reassembled from the initial stored template [7].

SRS is integrated with other devices to gain expected output, such as a regular microphone, or wireless video recording device. Along with these integrated devices, it can create inconsistent conditions and discrepancies in output. On the other side, inconsistent conditions arise from the physical and mental conditions of the user, such as age levels, illnesses, and depressions [8,9]; these could certainly trigger opportunities for security loopholes in the system. [1,3,10,11] are the observations that were done on implying the method of Mel Frequency Cepstral Coefficients (MFCC) in which they suggested further experiments using spectral centroid as the parameter that offers to resolve inconsistencies issues, despite devices integrated into the SRS itself, in SRS security, voice consistency is significant.

Speech signals are nonstationary and this fact contributes to the inconsistency issues. MFCC should be combined with another methodology that could handle nonstationary fact of signals. This paper surfaces a security resolution through in-house experiment pursuing consistency components to fit in security needs implying MFCC supported by Short Time Fourier Transform (STFT).

2. **Literature Review.** SRS is one of the biometric technology fields. In the biometric recognition methods survey, Delac and his team had stated that there are two category approaches in the voice recognition techniques generally which are Automatic Speaker Identification or ASI and Automatic Speaker Verification or ASV. The speaker identification attempts to use voices to identify an actual individual, while the speaker verification uses voices as the authentication attribute with more than one factor scenarios. In the first scenario of the speaker verification, a voice recognition process has to distinguish an individual by mapping his/her particular voice traits compared to what are stored in a database [12].

The mapping procedure shows the necessity of having voice systems to be trained for dealing with the time enrolments of individual's voice. The time enrolments are generally proceeded by feature extractions which typically deliver formants or sound characteristics uniqueness of each person's vocal tract. Individual voice traits serve to biometric characteristics for being distinctive and permanence at any given range of time. The characteristics of human's speech are invariant for an individual, but part of the feature pattern changes over time due to age, medical conditions and emotional state [4,12].

In biometric technology, there is a matching module to locate feature vectors to be computed and compared to stored data in the template. Fundamentally, it uses similar pattern of matching algorithms for voice recognitions and face recognitions. Then the decision-making module establishes the acceptance or rejection of the result of the matching module.

This establishment has proved that biometric technology with its implications of human physiological elements can perform as access keys to control a system [4,6,12].

There are some issues of limitations within biometric system which affect the data collectability and consistency faults in extent. In generating a template to be stored in database, biometric data collections from an individual during one-time enrolment to the next at the authentication process may have some differences which cause discrepancies for the sensor to interact correctly. The biometric traits eventually may be larger than what is expected and its similarity features cause the sensor to interact and capture the distinctiveness incorrectly. In some cases, there are speech recognition devices that do not necessarily require the biometric traits procedure giving any other users to have the access into the device. Furthermore, there is an attempt of forgery to the biometric traits with spoofing attacks which create disruptions or interferences and get through the access of the system [6,7,11,12]. Usually, the spoofing becomes easy to go through devices with the identification of voices. Figure 1 is a diagram to show how the biometric system gets spoofed.

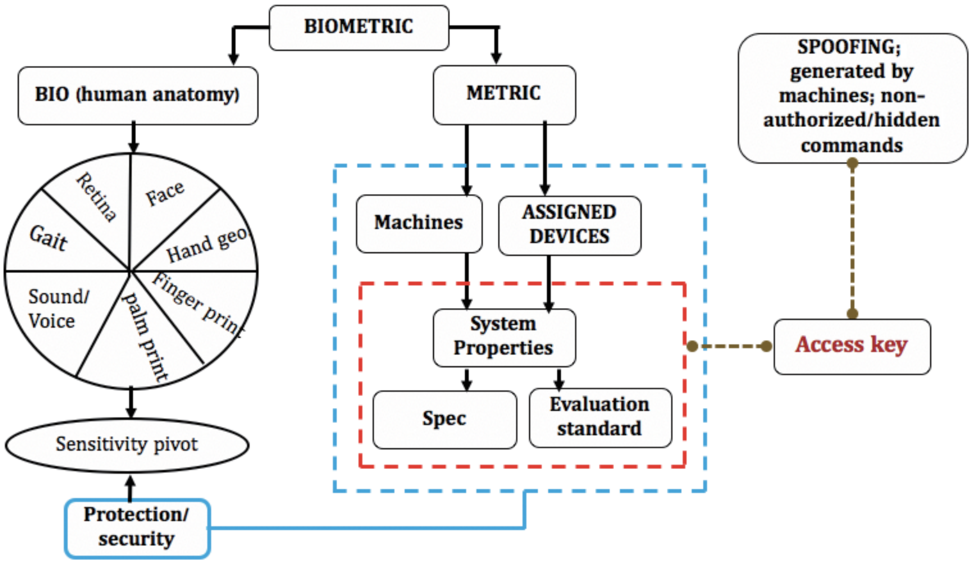

Figure 1. Spoofing in biometric system

Some approaches were applied in some databases by pursuing accuracies as resolutions in the biometric security. For example, in implying Subordinate Database, the Non-Negative Matrix (NNM) of 94% accuracy achievement is claimed to be better than 88% with the Hidden Markov Model (HMM). However, both NNM and HMM methods did not perform good when they were implied into TIMIT database. In the other case, HMM and ANN combined are used frequently in classification method. HMM seems like to have disadvantages that need to be covered by ANN to get better efficiency [13,14,17]. Here, it proves that the acquisition devices and its ambient or environment vary significantly which degrade the accuracy rates. Table 1 displays major methodologies frequently used.

Table 1. Methods and their techniques

Method's name	**Techniques**	**Remarks**
MFCC	Feature Extractions	Non-Linear
LPC (Linear Predictive Coeff.)	Feature Extractions	Linear
ANN (Artificial Neural Networks)	Classifications	Efficiency Performances
Hidden Markov Model	Classifications	88% accuracy
DTW (Dynamic Time Warping)	Classifications	Time Series
NNM (Non-Negative Matrix)	Classifications	94% accuracy
GMM (Gaussian Mixture Model)	Deep Learning	20% error reduction rate
SVM (Support Vector Machine)	Deep Learning	Efficiency Performances
DNN (Deep Neural Network)	Deep Learning	±20% error reduction rate
Neural Network	Deep Learning	Hardware architecture
Transformer	Deep Learning	Recent development

Despite some techniques in SRS, the fact is the need for a framework of extracted components to be done with its computations from combined methods to get final output [15]. The combination methods should be the most appropriate to fit the need of applications.

3. **Proposed Methodology.** Speeches are represented by sinusoid signals or complex exponentials that lead to issues such as formant and pitch period estimations, and applying analysis methods into the signal insight itself [16]. Those facts cause signals very vulnerable to a large sphere of variability, and the analysis becomes obscured since the same speaker does not always speak the same way and style in pronouncing the same word. Moreover, the variety of recording devices and transmission processes add to extend another complexity of the issues.

The graph below shows how a speech production configuration has to perceive that one speaking style which is significantly different in time manners [5,16]. Figure 2 displays the oscillations of sound waves or speeches. Figure 3 displays to feature those sound waves with their mathematic labels to represent amplitude (τ), period ($To = 2\pi/\omega o$), and time length (t). The equation labels listed here are included merely to support discussions for the following equations, and any elements that are out of scope within this paper should be disregarded. For example, the scope of this paper does not include imaginary component discussions.

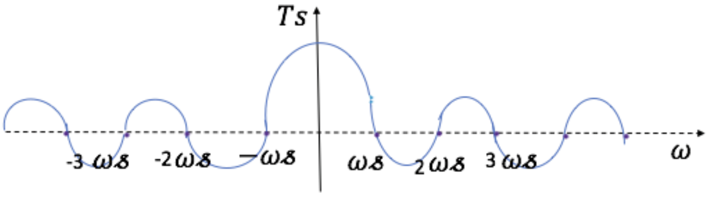

Figure 2. Oscillation of sound waves

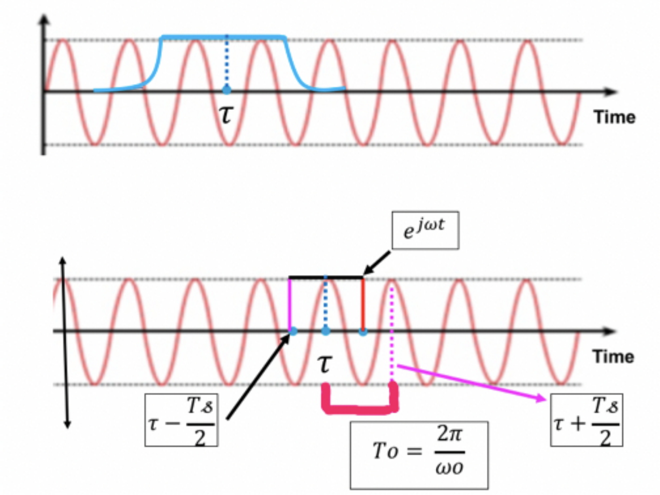

Figure 3. Signals and its periodical symbols

3.1. **MFCC.** In house experiment use Python programming language to process overall input speech signals into final output. MFCC is the deep learning methodology to imply extracting elements or parameters such as 'rmse', 'centroid', 'bandwidth', 'roll-off', 'zero crossing rate', and 'chroma'. However, for this paper scope, 'centroid' is captured as focal parameter discussions. Bibliography references will guide to more details for the rest of those terms.

Some audio features have been successfully applied to audio classification which are MFCC, LPC (Linear Predictive Coefficient), and LDB (Local Discriminant Bases) [17]. This paper is reinvigorating MFCC with its extracted parameter of spectral centroid that for some reasons had been overlooked, and utilizing STFT to mark them as security authentications.

Naturally, human's anatomy structures of vocal tract such as throat, tongue, or teeth rule how the speech sounds are generated and filtered. They should give receptions of speech sounds to other humans within receivable accuracies to depict the speech phonemes. The reception of speech sounds gives vibrations at some spots at human organ in the ear or cochlea, and conditionally those vibrations occur on frequencies entering along with speech sounds [18].

Speech sounds are formed or called as audio signals, and they are dynamic or persistently reshaping. Due to this condition, in controlling speech sounds, it is necessary to resolve them as stationary signals by having them in short time segments. The purpose of having them in much shorter signal (or within frames of 20 to 40 ms) is to obtain a reliable spectral estimate with assumptions of more stationed signals or less changes within those time frames. Each frame has a power spectrum and it identifies certain frequencies in the frame. Beside frequencies, these frames or spectral periodograms contain other information and involve some amounts of energy within each frame [19]. At this point, the need of extractions process is necessary.

MFCC is a system based on an auditory system that has synchronization with humans hearing [17,18]. In a sense, the auditory system is based on standard human perception so that normal sounds can be heard by the average human. The fact is that the frequencies in human auditory perception cannot keep up with the frequency of a sound that continues to scale linearly, or that the sound increases steadily from the lowest to the highest possible point. Each tone (symbolized in t) with the actual frequency is measured in Hertz (symbolized in Hz), and the level of the pitch point is individual or subjective with the Mel Scale measurement. Linearly, the Mel Scale frequency is spread at intervals below 1000 Hz and above 1 kHz in unit space with logarithmic values [11,17,18].

Features extraction is the initial important stage to identify linguistic contents of audio signal elements. The stage of feature extractions works as a repudiator for any cast-off audio signal due to speaker's background noises. The cochlea somehow could not distinguish frequencies that come through very nearly numbers due to various energy restraint in some frequency regions. In MFCC, this is performed by Mel filter bank: the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. Normally, human hearing cannot manage a linear scale of loudness, which means there must be huge energy variations needed to hear all levels of loudness [20].

As the frequencies get higher Mel filter banks get wider along to become less concerned about variations. This gives a person the ability of using Mel Scale to set filter banks energies space and width, and obtains their logarithms. This conditional operation closely leads to human hearing receptions, since the logarithm allows to use cepstral mean subtraction, which is a channel normalization technique. Human hearings are more comfortable to manage short variety of low-level frequency changes, and Mel Scale incorporates a pure tone of perceivable frequency to human hearings. However, those logarithm allowances cause overlapping filter banks while they are correlated to each other. The log filter bank energies need to get decorrelated to be used to classify features as certain models, and the Discrete Cosine Transform (DCT) computation helps to accomplish decorrelations of the energies. This DCT creates about 12 to 26 coefficients that represent filter bank energy variations which indicate SRS performances [19,20].

In the Mel Scale, every 1000 Mel is defined as a 1 kHz tone, with 40 dB above the perception of normal human hearing [5,6,11]. Audio classification methods generally involve extractions for distinguishing features from processing audio data into input data for classification model. Different approaches and various audio features were proposed with different success performances. In fact, those features can be extracted directly from the signal domain or from a transformation domain depending on the analytical approach used.

3.2. **Short Time Fourier Transform.** Short Time Fourier Transform (STFT) is a transformation method which is the development of the Fast Fourier Transform (FFT), and it certainly has significant reasons in finding a solution to a signal issue [21,23]. The FFT is a Fourier transform algorithm as an innovation of the Discrete Transform Fourier (DFT) to increase the speed of the computational process of the Fourier transform, or in other words, the application of the FFT algorithm reduces the looping process in the DFT [20,21,23].

As previously described, speech is a continuous time-varying signal with variations in syllable rate changes approximately every 10 times/second in intervals of 10 to 30 seconds [23]. From the length of those continuous signals, to get short duration segments as desired while simultaneously obtaining the computational results of the Fourier transform, then STFT is a method that has offered a solution for this. In the computational process, as shown at Equation (1), STFT regularly multiplies the duration of the longer time function $x[n]$ with the window function of the shorter duration $\omega[n]$. So the extraction process can be carried out on the desired segment without further modification [21,23].

$$\bar{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i \cdot 2\pi \cdot n \cdot \frac{k}{N}} \tag{1}$$

In other words, STFT offers an advantage of being able to transform non-stationary signals into a stationary signal representation by inserting a window function, shown at

Equation (2). In this case, the existing signal is extracted into several segments where those segments are assumed to contain stationary signals. In STFT there is a method called the Hamming window as an application that can display the frequency domain with smoother results [22,23]. A window function process could be either finite or infinite in duration principally, and in some cases have exponential windows computations, as shown at Equations (1) and (2).

$$X(\omega) = \mathcal{F}[x(t)] = \int_{-\infty}^{\infty} x(t) \cdot e^{-j\omega t} \partial t \tag{2}$$

STFT produces a time-frequency representation of a signal. A function of time and frequency is represented by $X[n,k]$ as shown at Equation (3). The variable n indicates the analysis window location along the time axis. The variable k denotes an index of frequency.

Here, unlike the general form of Fourier Transform, STFT can take the form of a normal distribution of a window function with the following formula shown in Equations (3) and (4) [23,24].

$$X[n,k] = \omega[n-m]x[m]e^{-j2\pi km/N} \tag{3}$$

$$m = n - (N-1) \tag{4}$$

In the term of frequency resolution, actually, there is no such an issue on FT using exponential kernel along the time. On the other hand, FT has a standard of building blocks in a complex exponential that oscillate over repeatedly along the time between -1 and $+1$. This affects FT to carry on having difficulties to represent signals that need to be localized in certain time as required. Those difficulties of localized functions span the frequency range, and this fact reveals uncertainty principles of FT. Meanwhile, STFT in particular time interval uses window kernel and has to define the infinity width of a window which in some points show nothing special about STFT [21,23,24]. Though FT is capable to identify time-domain sinusoidal components, STFT offers as a breakthrough those uncertainty principles by implying Fourier analysis to gain the achievable time and frequency resolution, as shown at Equation (5).

$$\text{STFT of } (t) = \cos(\omega \circ t)\frac{1}{2}\left[\left(e^{j\omega ot}\right)\right] + \frac{1}{2}\left[e^{-j\omega ot}\right] \tag{5}$$

This paper displays how STFT complements the MFCC to accomplish the overall objective of the conducted experiment. Figure 4 displays MFCC stages to STFT in mathematical equations flow diagram and Figure 5 is the overall display of methodology in a framework flow diagram.

The measurable input signal is defined to take speech signal and not others which mean to eliminate immeasurable segments such as speech pauses, silences, and any sounds that are not in words or speeches. Time series in FFT is needed to allocate selected speech segments, and it is visualized with the Mel Scale measurement results in its time and frequency domain. The STFT complements on eliminating those immeasurable segments of speeches [23,24]. Furthermore, in proceeding this framework, validation process is conducted to measure the data loss/errors.

4. **Proposed Security Procedures and Model.**

Prior to this paper, the experiment was conducted on observing the issues of age, gender, illness, and ambient noise factors exist as speech voice changes in speech recognition systems. Despite those issues, the experiment goal was to obtain the elements of voice features to secure the speaker's identity. MFCC was used to conduct observations on spectral centroid and bandwidth with data split into three groups: under 13, range 13 to 19, and above 19 years old for both male and female. The selected datasets are applied to the experiment objective in

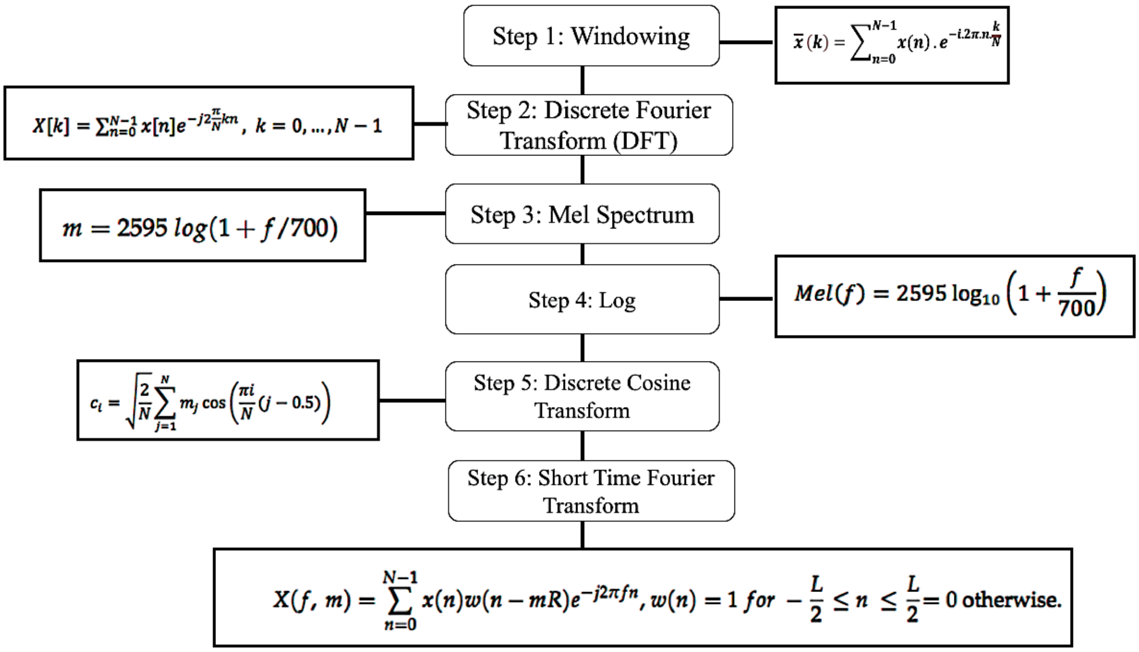
Step 1: Windowing
$\bar{x}(k) = \sum_{n=0}^{N-1} x(n) . e^{-i.2\pi.n.\frac{k}{N}}$
Step 2: Discrete Fourier Transform (DFT)
$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}, k = 0, \dots, N-1$
Step 3: Mel Spectrum
$m = 2595 \, log(1 + f/700)$
Step 4: Log
$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$
Step 5: Discrete Cosine Transform
$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$
Step 6: Short Time Fourier Transform
$X(f, m) = \sum_{n=0}^{N-1} x(n) w(n - mR) e^{-j2\pi fn}, w(n) = 1 \; for \; -\frac{L}{2} \leq n \leq \frac{L}{2} = 0$ otherwise.

FIGURE 4. MFCC to STFT mathematic equations flow diagram

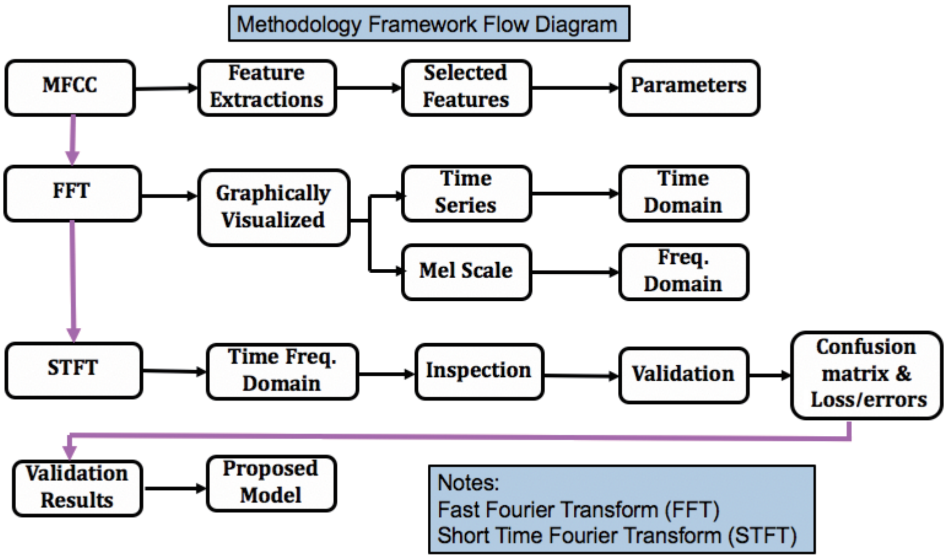

FIGURE 5. Methodology framework flow diagram

TABLE 2. Spectral centroid in age group

Age group	Under 13 years old	13 to 19 years old	Above 19 years old
Male	Centroid $\geq$ Bandwidth	Centroid $\geq$ Bandwidth	Centroid $<$ Bandwidth
Female	Centroid $>$ Bandwidth	Centroid $<$ Bandwidth	Centroid $\geq$ Bandwidth

finding how age distances (5 to 10 years distance), gender differences, and ambient noise will affect the speech voice changes. After applying the pitch test, stretch test, and noise adoptions on every group of selected datasets, the results conclude as Table 2.

In Table 2 observation results of noises adoption test in the system gave some situations such as spectral numeric was above the bandwidth, human beings naturally increase their speech voices when noises surrounded close to the area. This shows that spectral centroid has capacities to confront noises [19,26,29]. The objective of stretch test is to evaluate bandwidth intensity setting, and the results showed that bandwidth setting is acceptable due to the intensity followed provided bandwidth limit. The pitch test was conducted to evaluate maximum high and low tone for any audio input, the results showed that spectral bandwidth accommodates spectral centroid, this concluded that MFCC frequency setting is right. Moreover, the finding proves that MFCC effectively performed in capturing features of individual speech signal, and they can be adopted as critical elements for security parameters. The prior experiment had displayed comparison rates of different age levels and genders [11,26].

This paper is intended to extend the adopted parameters with a new focus on searching for consistency or stability speech voice in one individual dataset with 5 to 10 or more years distances. The gap in years is used to obtain repeated speech voices with similar or close range spectral centroid numeric. The objective is to show that despite aging, individual human beings have speech voice uniqueness to seal his/her identity unless affected by extreme illnesses. The uniqueness seals can be considered inclusively as personal security authentications.

Figure 6 shows overall research concept flow diagram which starts from dataset input to MFCC and Fourier series extraction processes and Deep Learning (DL) classifications. Epoch graph was used as a metric standard and validation procedure of proposed model, while data testing and verification are inclusively resumed here. The scaling stage is to apply iterations on extractions and classifications. The reason of having DLC meets the objective of the experiment itself which claimed to reach stability or consistency along the iterations process.

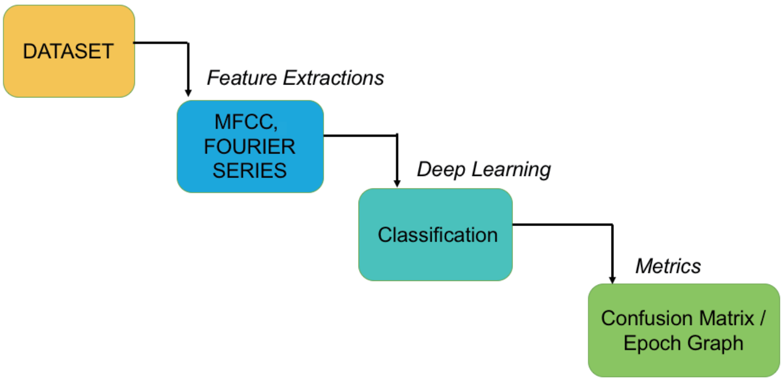

Figure 6. Overall research concept flow diagram

DL is a sub-component of Machine Learning (ML) with a computer file that had been trained by data scientists to accomplish assigned tasks within designed algorithm to obtain predictions and data analysis. ML works to reach accuracies, while DL works to deal with unstructured big data processing or having ambiguity issues. Regarding classification method in DL is to gain estimation values from finite set such as $\{1, \ldots, c\}$, and labelling each input signal that are formed from composed pairs which are called as a training set. This training set is arranged to reach parameter optimizations, so proposed model can be determined. Then the testing set is needed to measure estimation performance of the training set, and this overall operational procedure in DL is called as training protocol. The validation set evaluates and "mix-matching" those training and testing set to reach

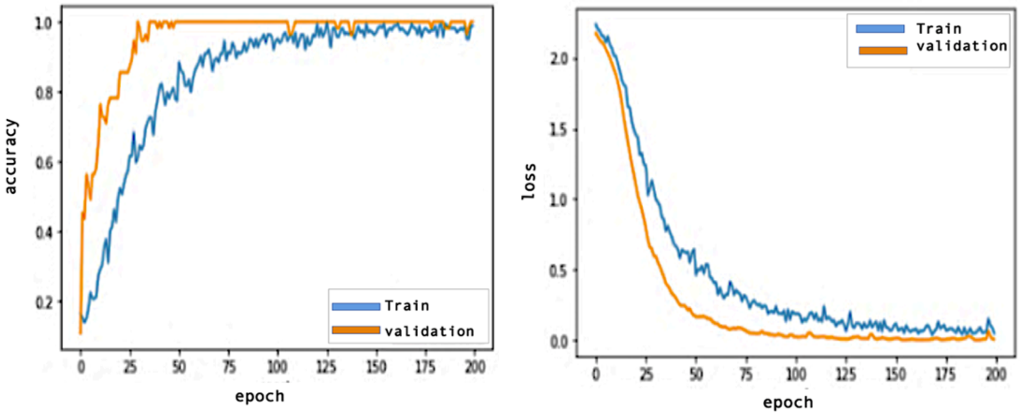

FIGURE 7. Left hand side: Training & Validation accuracy; Right hand side: Training & Validation loss

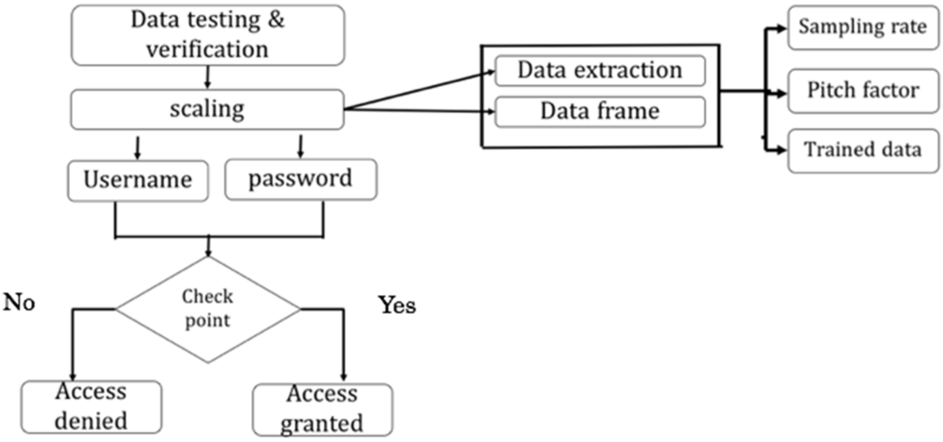

FIGURE 8. Proposed security flow diagram

the best configuration. Initially, the full training is conducted in first round, and the result is called epoch that declares dynamic losses which mean that downgrading values occur along the optimization process and this is called as a validation loss (Figure 7). This means its values reached at minimum level after having a certain amount of epoch and cause them to make increasing turn [30].

Figure 8 shows a proposed security procedure flow diagram that is applicable generally. If one of suggested speech voices options is selected as username or password, then it flows to the checkpoint procedure to be verified and saved in the device secured memory space. The device control access will proceed checkpoint verifications and administer to positive or negative confirmations.

5. **Dataset and Experiment Results.** The main objective of the experiment is to obtain individual speech voice character or uniqueness, despite aging factors. Datasets were distributed and selected on gender, ages in five to ten gap years. The gap of five to ten years audio dataset is considered to expose any selected dataset objected to pursue consistency elements. The intended consistency was expected to gain through exploration of speech recorded at intervals time of at least five to ten years which reflect with a user increasing years of age. Here, secondary dataset of year of 2000, 2008, and 2021 are presumed to fit the experiment objective. The selected dataset is meant to track the speech features along the time interval from the same individual and the dataset selection

was freed from conflict of interest in races, politics or religions. Following are the URL used as dataset:

https://youtube.com/_60-DgvRSdY
https://www.youtube.com/watch?v=pzBrKT0QLWc
https://www.youtube.com/watch?v=zSm1FbPcwZM

In reference to Figure 7 above, the graph of training and validation accuracy suggest that the applied model reaches the expected maximum point of 10 accuracies on training and validation, while the graph of training and validation loss shows that the applied model reaches its best after 200 epochs.

Figure 9 is the year 2000 dataset, and its output in time and frequency domain. Figure 10 is the year 2008 dataset, and its output in time and frequency domain. Figure 11 is the year 2021 dataset, and its output in time and frequency domain. The time gap distance has been designed with five to ten years or more to fit the objective of the experiment.

Total data drawn using Python is 23 spectral centroids data in 230 seconds for dataset for year 2000, 24 spectral centroids data in 240 seconds for dataset for year 2008, and 24 spectral centroids in 240 seconds for dataset for year 2021. Spectral centroid numeric was basically obtained from stated years to represent age levels of the speaker in overall 21 years as time length. The whole extracted dataset is expected to spotlight consistency

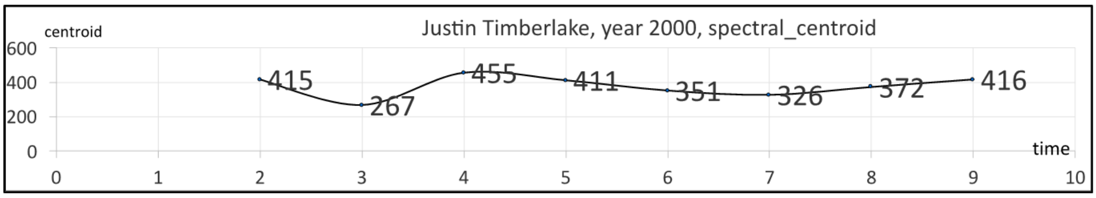

FIGURE 9. Extraction of dataset for the year 2000

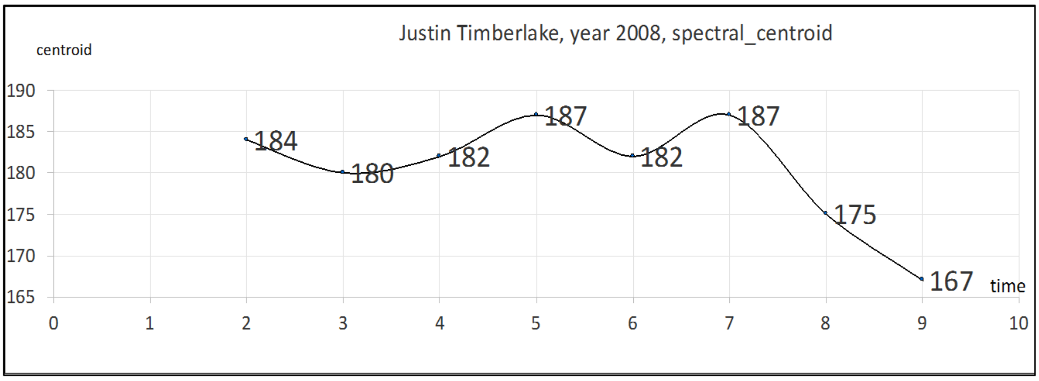

FIGURE 10. Extraction of dataset for the year 2008

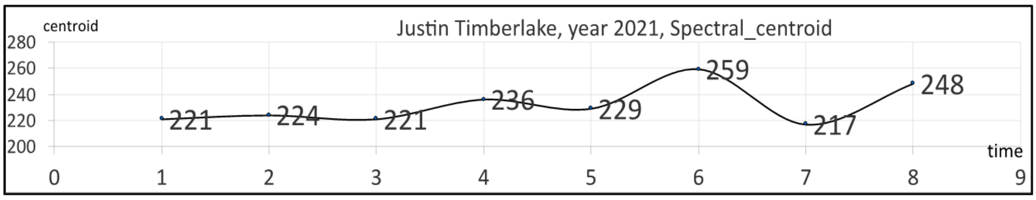

FIGURE 11. Extraction of dataset for the year 2021

on any interval time location. Dataset of the year 2000 and 2021 have almost similar situations; it contains the casual interview sessions. While the dataset of the year 2008 has a different situation from year 2000 and 2021, the 2008 was recorded directly from a microphone at the stage during half hour break of a football game.

Figure 12 displays the proposed schematic model into pursuing and gaining speech options used for login keys. Label 'user A' refers to one individual that has dataset year 1, year 2, and year n as input. Methodology processes data into expected features as output. Selected words or speeches are available for user A to have them as specific options in creating username or passwords implied into SRS devices.

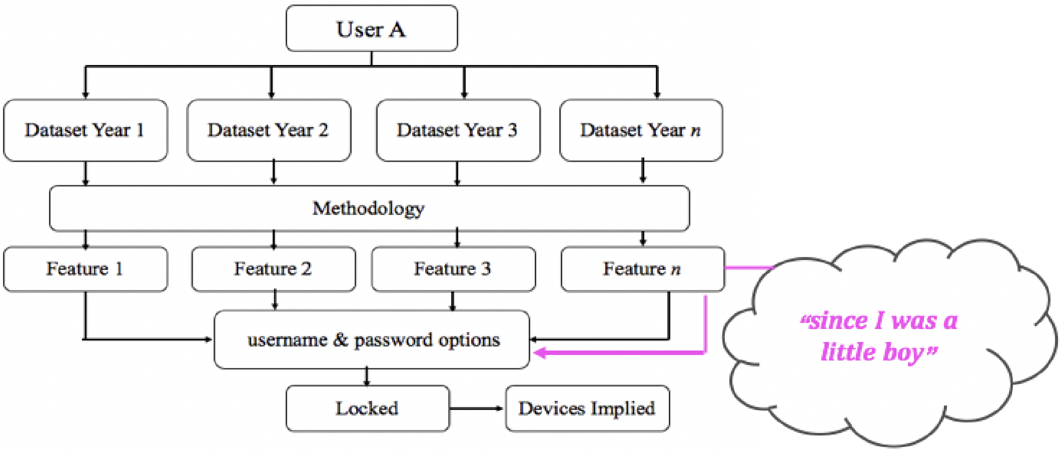

Figure 12. Applying suggested username/password as a login key

6. **Discussion and Analysis.** The reason of being absolute for spectral centroid supports the uniqueness of the individual's speech character [10-12] and it is exhibited by the experiment results. The proposed methods of MFCC and STFT are exhibited as supportive combined methods to proceed the output as the objective in pursuing the consistency in individual's speech despite age levels [10,11].

MFCC has been used as the methodology to generate its coefficients and other elements including spectral measurements of centroid and bandwidth contained in the audio dataset. Some research with its variety objectives has supported significant reasons that spectral centroid could be considered as decisive acoustic properties [11,12,19,26]. This paper displays the complement of STFT into MFCC for resolving the non-stationary facts of voice signals to succeed extraction processes. Table 3 shows that the extended efforts of having combined MFCC with other methods are revealing that MFCC remains significant to be reinvigorated [15,26,27]. MFCC is effective when combined with other methods to reach particular research objectives [19,20,29,31-39]. The table emphasizes on utilizing MFCC features to serve other methods for any objectives to gain. Comparative studies related to speech recognitions have tendencies to reach accuracies and optimizations. This paper is concentrated on finding unique features and sealing them as security features which are beyond accuracies and optimizations (or as a trade off in some cases).

Resuming Table 3 descriptions, here are some of them: Khamlich et al. conducted performance evaluation with MFCC implementation combined with Support Vector Machine (SVM) and Artificial Neural Networks of Multi-Layer Perceptron (MLP), their experiment was to pursue a comparative study of the performance algorithms applied to male and female speakers of different ages [20]. In the other side, Shantakumar et al. had MFCC combined with Dual Tree Complex Wavelet Transform (DTCWT) and Quick Fourier

Table 3. MFCC combined with other methods

No.	Combined MFCC methods	Classifiers	References
1.	MFCC, SVM, MLP	Self-data set	S. Khamlich et al., 2021
2.	DTCWT, MFCC, QFT	Self-data set	Shanthakumar et al., 2021
3.	MFCC, ENTROCY	Self-data set	Mohammed et al., 2021
4.	Discrete Wavelet, MFCC	Self-data set	Naing et al., 2019
5.	MFCC, DTW	Self-data set	Birch et al., 2021
6.	MFCC, Random Forest KNN, GMM	Gaussian Mixture Model (GMM)	Wicaksana and Zahra, 2021
7.	MFCC, LPC, CNN	GMM	Chowdhury and Ross, 2020
8.	MFCC, VQ	Self-data set	Gowda et al., 2019
9.	MFCC, BNN	Self-data set	Lubis and Gondawijaya, 2019
10.	MFCC, Baum Welch	Self-data set	Maseri and Mamat, 2020

Transform (QFT) to conduct an experiment to pursue performance evolution of face and SRS [29]. Birch et al., in another side, used MFCC and combined it with Dynamic Time Warping (DTW) to accomplish a comparative spoken words to user-dependent dictionary and reported its accuracies [33]. For advanced reading, the listed items in Table 3 encourage audiences to refer the attached bibliography lists.

Experiment results display data extractions of the year 2000 and 2021 which have similar context of interview scenes or settings between two persons. There are two voices recorded with some music backgrounds during the interview sessions. Since the observation objective pursues similar persons, the next challenging and critical matter is to separate which speeches are derived from the same speaker. These situations enable the observer to perceive in particular ways what MFCC offers to the human perception hearings [15,16,21,22,26]. At this point, observations are acknowledged to distinguish which voices or speeches are derived from interviewer or interviewee, such as familiarity of the person voice.

The observations start capturing the interviewee speech characters at the time when the object speaker has stability manner. For this case, Figure 9 has showed that spectral centroid of year 2000 dataset at the interval time location [13;2504] supports the words “completely me”, and the interval [20;3977] supports the words “eating me alive”.

Something interesting about finding spectral centroid from the experiment could surprise the observer. There are two interval time locations found with repeated words of “since I was a little boy” which is at interval time location [10;3804] and [20;3204]. Here, the spectral centroid numeric of 3804 and 3204 almost reach similar numeric patterns at different time stamps of 10 and 20. This could be considered as individual uniqueness that were gained from his vocal or speech tract characteristics, and this is what could be called as ‘options speech’ for login keys as displayed at Figure 5. Those spectral centroid numeric patterns in its slight difference could reach similar numeric pattern, when it is implied using the same hardware.

Figures 11 showed that the proposed concept could have been accomplished perfectly when those repeated words, with nearly close numeric patterns, are also found at year of 2008 or 2021 dataset, or 5 to 10 years gap as designed. However, dataset in this case is not supportive enough to fit in the design. Fundamentally, the proposed concept design is objected to have primary dataset instead of secondary used in the experiments. Recorded dataset should be within a periodical time range of 5 to 10 years to capture and save some similar typical words’ expression from the user. On the other hand, the ideal condition to the objective of proposed concept is to have a stable permanent hardware device to record user speeches.

Furthermore, the year 2021 dataset at interval time location [17;4207] supports the words of "final vocal", while the next other intervals were not captured with any similar words due to such interruptions by the interviewers or noise backgrounds. Moreover, the year of 2008, spectral data would not be sophisticated for pursuing speech characters of the speaker which is indicated from disruptions at the graph at Figures 10 and 11.

This is understandable due to the time and location of the dataset obtained in which the input instrument was the stage microphone at a football game during half time break. It is a convincing manner to know the high-end devices were supposed to be provided for this important football game to achieve a successful event; however, due to the experiment objective, unstable emotional speeches with ups and downs excitements during the event need to be ignored.

The SRS security access control brings pros and cons. Due to expansion of the Internet networks technology, one undeniable thing about SRS is excelling with its operational manner, where the user does not have to physically be in the location where access must be made. On the other side, it creates a loophole such as one case of remotely accessing sensitive data via telephone networks or even conventional computer networks. Despite all the Internet related technologies, and the origin motivations of pursuing security procedures with biometric system, the critical challenge is to have a biometric sensor in the same location with its users respectively. For example, the location of the microphone for speech processing must be available at the speaker's place instead of using a telephone.

The experiment was conducted using secondary dataset, with 5 to 10 years gap, which proved that the proposed model works. As stated in the end of the paper the experimental weakness in this case is simply dataset availability.

7. **Conclusions.** Security authentications could be tagged on with more advanced observations insightfully beyond the conventional procedures. The security concept in this paper revealed an insight view that the required consistency in security systems remains significant and critical. In biometric speech technology, the consistency is achievable despite the age level or stage changes. The finding of experiment processes in this paper displays one of the achievability ways to gain the required consistency. The proposed concept in this paper could be better implied with primary dataset and embedded in a robust hardware complemented with more methods combined which will encourage researchers in the future.

REFERENCES

[1] X. Li and M. Mills, Vocal features: From voice identification to speech recognition by machine, *Technology and Culture*, vol.60, no.2, pp.S129-S160, 2019.

[2] T. Sabhanayagam, V. P. Venkatesan and K. Senthamaraikannan, A comprehensive survey on various biometric systems, *International Journal of Applied Engineering Research*, vol.13, no.5, 2018.

[3] A. Y. Vadwala, K. A. Suthar, Y. A. Karmakar and N. Pandya, Survey paper on different speech recognition algorithm: Challenges and techniques, *International Journal of Computer Applications*, vol.175, no.1, 2017.

[4] Y. Chen, X. Yuan, A. Wang, K. Chen, S. Zhang and H. Huang, Manipulating users' trust on Amazon Echo: Compromising smart home from outside, *EAI Endorsed Transactions on Security and Safety*, http://creativecommons.org/licenses/by/3.0/, 2020.

[5] R. Thangarajan and A. M. Natarajan, A robust front-end processor combining Mel frequency cepstral coefficient and sub-band spectral centroid histogram methods for automatic speech recognition, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol.2, no.2, 2009.

[6] T. Ignatenko and F. M. J. Willems, Biometric security from an information-theoretical perspective, *Foundations and Trends in Communications and Information Theory*, vol.7, no.2, 2010.

[7] T. Gunendradasan, B. Wickramasinghe, P. Ngoc Le, E. Ambikairajah and J. Epps, Detection of replay-spoofing attacks using frequency modulation features, *Proc. Interspeech*, pp.636-640, 2018.

[8] H. Beigi, *Fundamentals of Speaker Recognition*, Springer New York, NY, 2011.

[9] S. Feng, O. Kudin, B. M. Halpern and O. Scharenborg, Quantifying bias in automatic speech recognition, *arXiv Preprint*, arXiv: 2103.15122v2, 2021.

[10] L. P. Ngoc, A. Eliathamby, E. Julien, S. Vidhyasaharan and H. C. C. Eric, Investigation of spectral centroid features for cognitive load classification, *Speech Communication*, vol.53, no.4, pp.540-551, 2011.

[11] H. I. Pratiwi, I. H. Kartowisastro, B. Soewito and W. Budiharto, Adopting centroid and bandwidth to shape security line, *2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM)*, Laguboti, North Sumatra, Indonesia, 2022.

[12] K. Delac and M. Grgic, A survey of biometric recognition methods, *The 46th International Symposium Electronics in Marine*, Zadar, Croatia, 2004.

[13] L. H. Acosta and D. Reinhardta, A survey on privacy issues and solutions for voice-controlled digital assistants, *Pervasive and Mobile Computing*, vol.80, http://dx.doi.org/10.1016/j.pmcj.2021.101523, 2022.

[14] J. S. Edu, J. M. Such and G. Suarez-Tangli, Smart home personal assistants: A security and privacy review, *ACM Computer Survey*, vol.53, no.6, 2020.

[15] M. Malik, M. K. Malik, K. Mehmood and I. Makhdoom, Automatic speech recognition: A survey, *Multimedia Tools and Applications*, vol.80, pp.9411-9457, 2021.

[16] J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz and M. Brudno, Speaker attribution with voice profiles by graph-based semi-supervised learning, *arXiv Preprint*, arXiv: 2102.03634v1, 2021.

[17] P. P. Patange and J. S. R. Alex, Implementation of ANN based speech recognition system on an embedded board, *International Conference on Nextgen Electronic Technologies: Silicon to Software (IICNETS2)*, 2017.

[18] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas and J. Droppo, SynthASR: Unlocking synthetic data for speech recognition, *arXiv Preprint*, arXiv: 2106.07803v1, 2021.

[19] D. Prabakaran and S. Sriupilli, Speech processing: MFCC based feature extraction technique – An investigation, *Journal of Physics Conference Series*, vol.1717, 0120009, 2021.

[20] S. Khamlich, I. Atouf, F. Khamlich and M. Benrabh, Performance evaluation and implementations of MFCC, SVM and MLP algorithms in the FPGA board, *International Journal of Electrical and Computer Engineering Systems*, vol.12, no.3, 2021.

[21] H. Sujadi, Signal processing system using fast Fourier transform algorithm, *Proc. SINTAK*, 2017.

[22] M. M. Bachtiar, B. S. B. Dewantara and D. Prastyo, Home monitoring and control using smartphone and speech processing, *International Joint Conference on Science and Engineering (IJCSE 2020)*, vol.196, 2020.

[23] S. Meignen, D.-H. Pham and M. A. Colominas, On the use of short-time Fourier transform and synchrosqueezing-based demodulation for the retrieval of the modes of multicomponent signals, *Signal Processing*, vol.178, no.6, https://www.elsevier.com/open-access/userlicense/1.0/, 2020.

[24] K. Li, H. Rüdiger and T. Ziemssen, Spectral analysis of heart rate variability: Time window matters, *Front Neurol.*, doi: 10.3389/fneur.2019.00545, 2019.

[25] J. H. L. Hansena, On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks, *Speech Communication*, vol.101, pp.94-108, 2018.

[26] A. Madhavaraj, T. V. A. Padmanabha and A. G. Ramakrishnan, Subjective and objective experiments on the influence of speaker's gender on the unvoiced segments, *arXiv Preprint*, arXiv: 1807.05813v1, 2018.

[27] W. Mustikarini, R. Hidayat and A. Bejo, Real-time indonesian language speech recognition with MFCC algorithms and Python-based SVM, *IJITEE*, vol.3, no.2, 2019.

[28] D. Y. Mohammed, K. Al-Karawi and A. Aljuboori, Robust speaker verifications by combining MFCC and entrocy in noisy conditions, *Bulletin of Electrical Engineering and Informatics*, vol.10, no.4, pp.2310-2319, 2021.

[29] H. C. Shantakumar, G. S. Nagaraja and M. Basthikodi, Performance evolution of face and speech recognition system using DTCWT and MFCC features, *Turkish Journal of Computer and Mathematics Education*, vol.112, no.3, pp.395-404, 2021.

[30] F. Fleuret, *The Little Book of Deep Learning*, Universite De Geneve, https://fleuret.org/public/lbdl.pdf, Accessed and download on June 16, 2023.

[31] Y. M. Duraid, K. Al-Karawi and A. Aljuboori, Robust speaker verification by combining MFCC and entrocy in noisy conditions, *Bulletin of Electrical Engineering and Informatics*, vol.10, no.4, doi: 10.11591/eeiv10i4.2957, 2020, 2021.

[32] H. M. S. Naing, R. Hidayat, R. Hartanto and Y. Miyanaga, Discrete wavelet denoising into MFCC for noise suppressive in automatic speech recognition system, *International Journal of Intelligent Engineering & System*, 2019.

[33] B. Birch, C. A. Griffith and A. Morgan, Environmental effects on reliability and accuracy of MFCC based voice recognition for industrial human-robot-interaction, *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol.235, no.12, pp.1939-1948, 2021.

[34] V. S. Wicaksana and A. Zahra, Spoken language identification on local language using MFCC, random forest, KNN, and GMM, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.12, no.5, 2021.

[35] A. Chowdhury and A. Ross, Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals, *IEEE Transactions on Information Forensics and Security*, vol.15, 2020.

[36] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan and C.-T. Sun, Efficiently classifying lung sounds through depth wise separable CNN models with fused STFT and MFCC features, *Diagnostics*, vol.11, no.4, 732, https://doi.org/10.3390/diagnostics11040732, 2021.

[37] M. Maseri and M. Mamat, Performance analysis of implemented MFCC and HMM-based speech recognition system, *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Kota Kinabalu, Malaysia, 2020.

[38] C. Lubis and F. Gondawijaya, Heart sound diagnose system with BFCC, MFCC, and backpropagation neural network, *IOP Conference Series: Materials Science and Engineering*, vol.508, 012119, 2019.

[39] V. P. Gowda, M. Murugavelu and S. Thangamuthu, Recognition based on threshold using MFCC and VQ, *International Journal of Electrical and Computer Engineering (IJECE)*, vol.9, no.6, pp.4684-4695, 2019.

Author Biography

Heni Ispur Pratiwi obtained her Bachelor of Science degree in Computer Engineering and Technology of Arizona State University, United States, in 2006, and Master of Science in Electronic Engineering Technology at Arizona State University, United States, in 2008. Currently she is finalizing her doctoral program at Bina Nusantara University, Jakarta, Indonesia. Her research interests are in multi-agent security pattern recognitions.

Widodo Budiharto had obtained his B.Sc. in Physic Major from University of Indonesia, Master degree in Information Technology from STT Benarif, Jakarta, Indonesia, and Doctoral degree in Electronic Engineering of Sepuluh November Technology Institute, Surabaya, Indonesia. He worked as Visiting Professor at Erasmus Mundus French Indonesian Consortium (FICEM), France, Universitas Hosei, Japan, and Erasmus Mundus Scholar pada EU Universiteit de Bourgogne, France, in 2017, 2016, and 2007. Currently he posts as Academic Professor for artificial intelligence at School of Computer Science, Bina Nusantara University, Jakarta, Indonesia. His specific topic areas are in intelligent systems, data science, robot vision, and computational intelligence.

Iman Herwidiana Kartowisastro has been lecturing and conducting research at Bina Nusantara (BINUS) University, Jakarta, Indonesia since 1991. He graduated from Trisakti University, Jakarta, Indonesia in 1986 with major in Electronics and Telecommunications. He obtained both of his post graduate degrees from the City University, London, UK, which are M.Sc. in Information Engineering in 1987 and Ph.D. in Robotics Control in 1991. His research interests are in computer vision, computational intelligence, and signal processing.

Benfano Soewito had obtained his B.Sc. in Physic Major of MIPA Department at Airlangga University, Surabaya, Indonesia. His M.Sc. and Ph.D. degrees were obtained in 2004 to 2009 from Computer and Electrical Engineering Department, Southern Illinois University, United States. He started lecturing in Bakrie University since 2013, and currently he posts as a Professor for Graduate Programs of Bina Nusantara University, Jakarta. His research interests are in information technology specified in Internet packet processing and scanning, router development, security and computer network.