International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.14, No.1 (Oct-23)

http://dx.doi.org/10.12785/ijcds/140192

Syllable-based Speech Recognition System Using Pitch Detection on Time–Frequency Domain Feature Extraction

Sopon Wiriyarattanakul[1], Piroon Kaewfoongrungsi[2] and Ekkalak Sumonphan[3]

[1]*Program in Computer Science, Faculty of Science and Technology, Uttaradit Rajabhat University, Uttaradit, Thailand,*
[2]*Department of Computer, Faculty of Science and Technology, Chiang Mai Rajabhat University, Chiang Mai, Thailand*
[3]*Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology LANNA Tak, Thailand*

Received 3 Nov. 2022, Revised 18 Sep. 2023, Accepted 21 Sep. 2023, Published 1 Oct. 2023

Abstract: This research presents the segmentation of single-syllable sounds for speech recognition using an artificial neural network. The network combines key features from speech signals in the time and frequency domains. The approach involves dividing speech signals into frames using the short-time energy waveform. Pitch markers are then extracted from the frames and used as reference points to split them into sections. The sections are further analyzed using window searching to identify positions, amplitudes, local minimum and maximum values, and maximum slope values, which serve as key features in the time domain. In the frequency domain, cepstrum coefficients on the Mel scale are used as additional key features. The two types of key features are combined for speech recognition using the artificial neural network. The study also compares the performance of the combined and separated key features in the time and frequency domains when fed into the neural network. The results demonstrate that using the artificial neural network with two input layers (Mel frequency cepstral coefficient and time domain features) and the same hidden layers yields the highest recognition accuracy of 96.97% and 88.43% for blind tests.

Keywords: Pitch detection, Time-Frequency domain, Feature extraction, Speech recognition, Syllable, Short-time energy waveform

1. Introduction

Nowadays, computerized speech processing has gained significant importance in facilitating communication between humans and computers. The advancement of computerized speech recognition is a direct outcome of the continuous development in computer technology, encompassing both hardware and software components. Hardware capabilities have been continuously improving, with faster processing speeds, larger memory capacities, and reduced power consumption per command. Moreover, software innovations in pattern recognition learning and artificial intelligence, which aim to mimic human recognition abilities, have led to the widespread adoption of computers for speech signal processing across various applications. Speech recognition begins with an initial stage of signal processing, wherein audio recordings are converted into numerical signals to prepare the speech for further analysis. These numerical signals then serve as the speech data for subsequent processing steps. Typically, researchers employ fundamental processing principles, including filtering, to eliminate undesired frequencies using digital filters. Another crucial step is endpoint detection, which identifies the beginning and end points of speech by segmenting it into verbal and non-verbal sections, as well as individual speech frames. Essentially, endpoint detection involves dividing the speech signal into shorter segments for more focused analysis. The subsequent stage is a crucial step known as feature extraction. It entails analyzing the feature values present in the speech signals and deriving a set of coefficients that effectively represent the speech. This process is vital as it eliminates irrelevant components, allowing for accurate representation of the speech. By effectively extracting and deriving features, it becomes possible to distinguish between different speech signals, thereby achieving efficient outcomes in speech processing. There are two approaches to extracting feature values, the first of which involves extracting feature values in the time domain. This approach has found applications in various fields. For instance, in Dysarthria detection, researchers have observed that the jitter and shimmer features in the time domain of disordered voice samples tend to be higher compared to those of healthy voice samples [1] . Similarly, in speaker identification, time domain features have been utilized by calculating various statistics for each window size in the training set. These statistics include mean, median, mode, standard deviation, variance, covariance, zero cross rate, minimum, maximum, root mean square, and distance. By incorporating these 11 features into a numeric master feature vector, researchers achieved a remarkable accuracy of 96.9% for gender identification using a random forest classifier [2]. Additionally, a technique called time encoded signal processing and recognition (TESPAR) has been applied to speaker identification. TESPAR is an approximation method that analyzes and describes time-varying signals in a computationally efficient manner [3]. Another intriguing and challenging study focuses on the recognition of spoken English letters and leverages features in the time domain. In this study, the researchers utilized extracted features such as the number of threshold crossings, peak-to-peak amplitude, minimum, maximum, and average sampling. The results were promising, achieving a commendable recognition rate of 92.2%. This outcome serves as evidence for the potential of incorporating the proposed solution principle into contemporary automatic speech recognition systems[4]. As evident from the aforementioned examples, feature extraction in the time domain proves to be a valuable approach in speech recognition. It offers advantages such as reduced computation time and the ability to capture features that represent phase shifts in speech signals. However, it is worth noting that the second approach, which involves extracting feature values in the frequency domain, has gained more popularity. In speech recognition applications, feature extraction in the frequency domain is commonly utilized. One widely adopted technique is the Mel frequency cepstral coefficient (MFCC). In this technique, the Cepstrum is obtained by performing an Inverse Fourier Transformation on the logarithm of the sound signals' spectra over a short period of time. The MFCC approach enhances the cepstrum by adjusting the scales of the spectrum to align more closely with human hearing. As low-frequency speech signals tend to carry more important information than high-frequency signals, the spectrum scale is designed to capture finer details in the low-frequency range. This scaling is achieved using the Mel scale [5]. The feature extraction process, known for its effective frequency analysis in the speech range, has been widely implemented in various studies. For instance, it has been utilized in controlling a five degree of freedom robot arm with an Arduino microcontroller for tasks such as object pick and place [6]. Moreover, the application of principal component analysis to MFCC features has been employed to extract the most significant components for recognition in spoken word databases [7]. In another study, MFCC was applied to recognize numbers 0–9 in the Pashto language using the K-nearest neighbor algorithm [8]. Additionally, when comparing the performance of text dependent speaker recognition using MFCC features, linear predictive coding (LPC), and discrete wavelet transform (DWT), the results showed that MFCC outperformed LPC and DWT [9]. While the MFCC feature offers advantages in frequency-based analysis and finds applications in various domains, it is important to note its limitations. One notable weakness is observed when dividing the signal into fixed frame segments of around 20–40 milliseconds. In this approach, each frame may not align accurately with the signal's true periodicity. This discrepancy arises due to the fixed frame size, which fails to consider the signal's actual period. Consequently, the resulting frame lengths may not adequately represent periodic signals during the MFCC extraction process. Another concern is the issue of DFT leakage introduced during the initial step of calculating the Discrete Fourier Transform (DFT). However, this concern can be addressed by utilizing pitch-based segmentation [10]. It is also worth mentioning that the feature extraction using MFCC solely relies on the coefficients of the full Fourier transform, resulting in the loss of phase information in the signal. Unfortunately, the phase information of the speech signal is not taken into consideration in this method. Therefore, this research aims to showcase the utilization of feature extraction in the time domain, which allows for the inclusion of phase information through a simpler calculation process. This approach is then combined with spectral features obtained from MFCC, which provide specific frequency-related information about the signal. Moreover, a multilayer perceptron neural network is utilized as a tool for speech recognition, enabling the assessment of the effectiveness of incorporating significant features from various speech patterns in the recognition process.

2. Methodology

This section illustrates the research methodology adopted in the study, as depicted in Figure 1. The methodology comprises four distinct steps, which will be elaborated upon in the following sections.

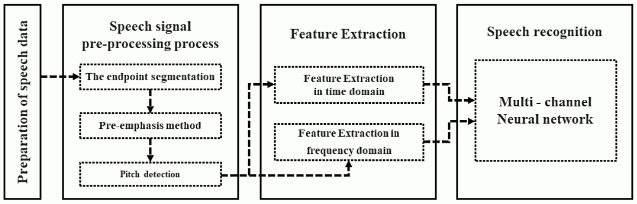

Preparation of speech data
Speech signal pre-processing process
The endpoint segmentation
Pre-emphasis method
Pitch detection
Feature Extraction
Feature Extraction in time domain
Feature Extraction in frequency domain
Speech recognition
Multi - channel Neural network

Figure 1. Syllable-based speech recognition using pitch detection on time–frequency domain feature extraction

A. Preparation of Speech Data

The speech audio data utilized in this study was obtained from [10], encompassing a group of 50 speakers that included both male and female participants. The training dataset comprised 11 different pronunciations of common words in the Thai language. These words were repeated 10 times by 20 male speakers and 20 female speakers, resulting in a total of 4,400 words. To facilitate the training process, the dataset was divided into 10 subsets, each containing 440 words obtained from each speaking session. For the blind test, a separate test dataset was created. This dataset included the same words as those in the training dataset and was repeated 10 times by 5 male speakers and 5 female speakers who were not part of the training dataset. This resulted in a total of 1,100 words for the blind test. To capture the speech signals, a sound card in a personal computer was utilized. The analog signals received through a microphone were converted into a digital format at a sampling rate of 22,050 samples per second with a resolution of 16 bits per sample. The recorded speech signals were then saved in .wav file format for further analysis and processing.

B. *Speech Signal Pre-processing Process*

Preliminary speech signal processing is a crucial step in preparing the speech audio data before extracting key characteristics of the speech signal. In this research, two important processes, namely endpoint segmentation and pitch detection, are employed for speech signal processing. Endpoint segmentation aims to identify the starting and ending points of the speech signal in order to isolate the speech area for further processing. The non-voice or unvoiced regions are discarded by removing the parts with lower energy values than a specified threshold. This helps to focus on the speech-specific regions while filtering out unwanted noise. The next step in pre-processing involves noise reduction in the time domain, which can be achieved using methods such as pre-emphasis or highpass filtering. This helps to minimize background noise and enhance the clarity of the speech signal. Pitch detection is another crucial aspect of the pre-processing stage. Pitch refers to the semi-repetitive pattern (quasi-periodic) that occurs in the waveform of speech signals. It is related to the frequency of speech and varies across different vowels and tones. Pitch detection involves estimating the reference position of the pitch marker, which represents the beginning of pitch periods, as well as determining the length of the pitch or the fundamental period of a periodic signal in speech signal processing. Pitch detection plays a significant role in various applications of speech signal processing, including speech recognition [11], [12], emotion recognition [13],[14], improving speech clarity [15],[16] and [17] and speech synthesis [18],[19]. There are different methods available for pitch detection, including the autocorrelation function (ACF) [20], YIN estimator [21], and short-time energy waveform (SEW) [22]. These methods are designed to locate the pitch position or estimate the fundamental frequency of the signal, and they can be applied in either the time domain or the frequency domain. In this research, the SEW method was selected for pitch detection due to its ability to accurately detect pitch and determine the optimal frame size for pitch detection in Figure 2.

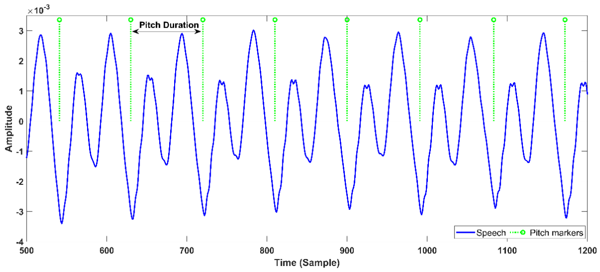


Figure 2. Position of pitch marker and pitch duration on speech signal.

After the endpoint segmentation process, which isolates the voice segments, the resulting voice signals were divided into fixed frame segments. Each frame had a duration of approximately 20–40 milliseconds. Within each frame, the voice signal was analyzed to determine the feature values. In this research, a pitch-based segmentation approach was employed to divide the speech signal frame into segments of varying sizes. The rationale behind using this approach was to mitigate the occurrence of Spectrum leakage problems during the calculation of DFT, which is a crucial step in determining the cepstrum coefficients on the Mel scale. Pitch-based segmentation also facilitates the manipulation of data in the time domain by utilizing the position of the pitch marker as a reference point for selecting and dividing the voice signal. Pitch detection using the SEW involves analyzing signal segments of the input signal x(n) with a length of N. It quantifies the energy of these signal segments as a function of time:

$$E(n) = \sum_{i=n}^{n+W-1} x^2(i) \tag{1}$$

Where W is a window size and n = 0,. . . , N–W.

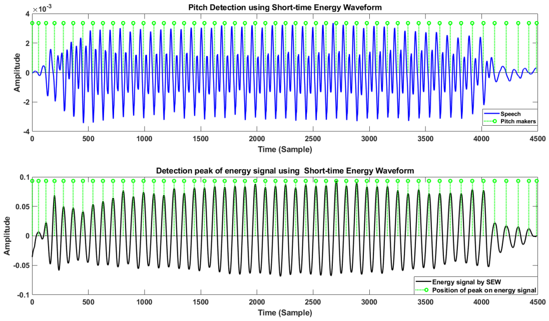


Figure 3. Pitch markers represent peak of energy signal, which corresponds to references in the speech signal.

The SEW process involves extracting the energy value of a speech signal, denoted as E(n), within a short time period at a specific moment. This process is particularly effective for pitch detection due to the characteristic waveform of the pitch, which alternates between low and high energy levels. Consequently, the peak values in the energy waveform can serve as reference points for identifying pitch markers within the voice signal. Figure 3 illustrates the derivation of the pitch marker and its utilization as a reference point for segmenting the voice signal into frames.

C. *Features of Speech Signals Determination*

In order to determine the features of speech signals for speech recognition, it is crucial to choose suitable data sets that accurately represent the speech signals. In this research, two approaches were employed to characterize the speech signals and extract their features. The first approach involved analyzing the signals in the time domain, while the second approach focused on characterizing them in the frequency domain.

1) Features Extraction in Time Domain: This research introduces an innovative method for extracting features from voice signals in the time domain. The approach focuses on capturing important information such as position, amplitude values, local minimums, local maximums, and max slopes. To achieve this, a window searching technique is employed, where a variable is created to store a dataset of size N, starting from the initial position of the voice signal. The collected dataset, consisting of amplitude values, is then analyzed to extract the relevant features. By examining the data, local minimums and local maximums can be identified as the lowest and highest values respectively, along with their corresponding positions. The window is then shifted progressively from left to right, one point at a time, until it reaches the end of the voice signal. This process enables the extraction of values for local minimums, local maximums, and their associated positions on the time axis (sample positions), as depicted in Figure 4. The size of the searching window directly influences the number of feature values obtained. A smaller searching window will generate a higher number of key characteristics, whereas a larger searching window will yield a lower number of feature values. In this study, a 30-point searching window was selected for analysis.

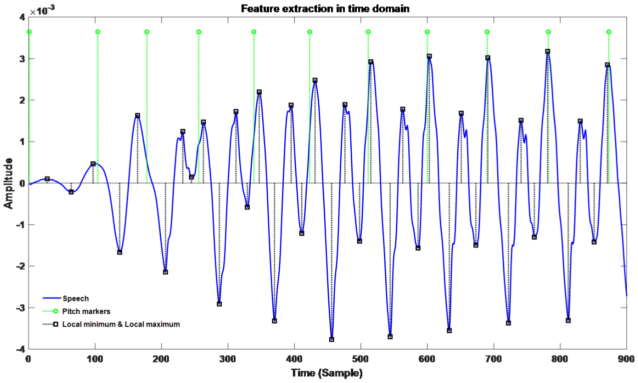


Figure 4. Positions of local minimum and local maximum using window searching

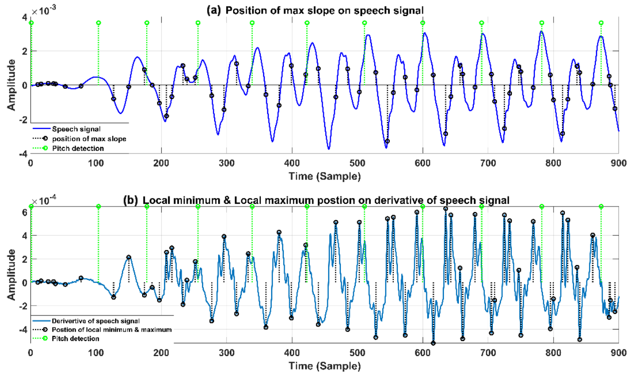


Figure 5. Relationship between max slopes, local minimums, and local maximums on the 1st derivative of the voice signal

Similarly, the max slope value can be obtained by applying window searching to the first derivative of the voice signal. By utilizing window searching on the first-order derivative signal, the position of the maximum slope in the voice signal can be determined. Window searching, as a direct calculation method for extracting important features in the time domain, allows for capturing phase change information in the signal that cannot be achieved using the frequency domain. Figure 5 (b) visually presents the positions of local minimums and local maximums on the first derivative of the voice signal in the time domain. It highlights the existence of a significantly steep midpoint located between the highest and lowest peaks of the voice signal, as shown in Figure 5 (a).

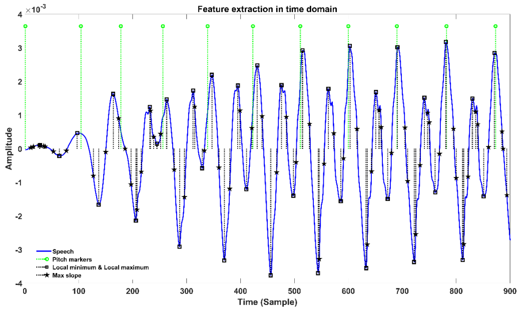


Figure 6. Local minimum and local maximum and max slope positions on the voice signal

Figure 6 depicts the positions of local minimums, local maximums, and max slopes, which are vital key attributes in the time domain. These features will be referred to as important features in subsequent discussions, as shown in Figure 7 (a). To provide a closer look at the time axis, Figure 7 (b) presents a magnified view ranging from 0 to 900 in the time domain. It clearly demonstrates that the positions of these features are distributed throughout one complete signal period, offering a comprehensive representation of the time domain.

Figure 7 (c) depicts the positions of features in the time domain for each pitch. It is noticeable that the number of points varies depending on the duration of each pitch. This variation is a result of the voice audio signal's periodic nature, which changes over time. Specifically, at the beginning of the voice, corresponding to the initial consonant sound, the voice signal exhibits significant fluctuations. In contrast, mid-tones and endings, which correspond to vowels and similar signal periods, exhibit a more consistent number of feature points. The discrepancy in the number of features obtained for each pitch duration has implications for defining the input in the neural network. Therefore, a process is needed to ensure an equal number of features is obtained for each pitch duration.

In this study, the positions of the three highest amplitudes and the three lowest amplitudes were selected within each pitch duration. The analysis considered not only the amplitude values but also the corresponding time values in the time domain (represented on the y-axis) to capture the changes in the signal period. This resulted in obtaining two parameters at each selected point: the amplitude value denoted as "A" and the time value denoted as "C". By selecting the three highest and three lowest amplitudes, a total of six amplitudes were obtained within one pitch duration. When combining these amplitudes with their respective time domain values, a total of 12 data parameters were acquired per pitch duration. Therefore, for a voice audio signal with 50 pitch durations and selecting six amplitude points per duration, the utilization of features in the time domain involves 600 parameters. These parameters consist of 300 "A" parameters (amplitude values) and 300 "C" parameters (time values).

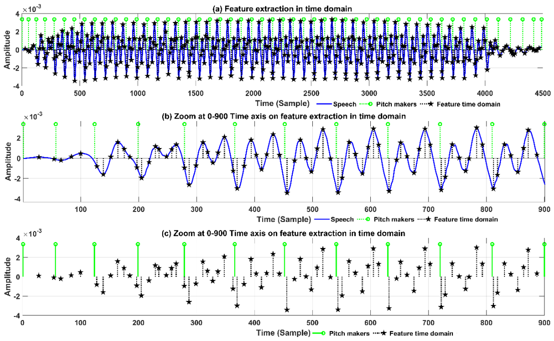
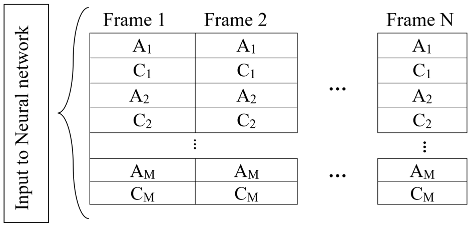


Figure 7. Positions of features in the time domain

a) Adjusting parameter C to the Same Range: One challenge in implementing parameter C (time axis value) is the variability in voice among different individuals. Even when speaking the same words, individuals have different fundamental frequencies, resulting in variations in the length of their speech pitches. As a result, the sets of parameter C for speech pitches with different fundamental frequencies may appear similar, despite representing different durations. To address this issue, it is necessary to scale parameter C in the time domain. Scaling parameter C ensures that it falls within a consistent range by utilizing a formula that estimates C based on the length of the specific pitch. This scaling approach takes into account the duration or fundamental frequency of the speech signal and adjusts parameter C accordingly. By applying this scaling technique, the parameter C values can be normalized across different individuals, allowing for more accurate and reliable comparisons and analysis of speech features.

$$C_Normalized = \frac{c}{Pitch\ length} \qquad (2)$$

b) Parameter Formatting of Features in the Time Domain: To utilize the parameters A and C in the neural network, they were arranged as N-dimensional vectors. In one frame, there were five pitch durations, each consisting of six parameters A and C. Therefore, one frame contained a total of 30 parameters A and C. The frames were structured to overlap with the previous frame for four pitch durations, as depicted in Figure 8.



Frame 1	Frame 2		Frame N
A_1	A_1		A_1
C_1	C_1	...	C_1
A_2	A_2		A_2
C_2	C_2		C_2
⋮			⋮
A_M	A_M	...	A_M
C_M	C_M		C_M

Figure 8. Parameter arrangement of features in the time domain as vectors that can be fed to neural networks.

2) Features Extraction in Frequency Domain by MFCC: The study on preventing DFT leakage [10] has shown that can arise when dividing the signal into fixed frame segments of the same size. To mitigate this problem, the pitch-based segmentation method was applied in this study for determining the cepstrum coefficients on the Mel scale (MFCC). In this research, the speech signals were framed using a frame size equivalent to 5 contiguous pitch durations, which approximates the typical framing duration of 20–40 milliseconds. Figure 9 illustrates the position of the pitch markers, indicating how the speech signal was divided into frames. Each frame overlapped with the previous frame for 4 pitch durations, resulting in a total of 44 frames for the analyzed speech signal.

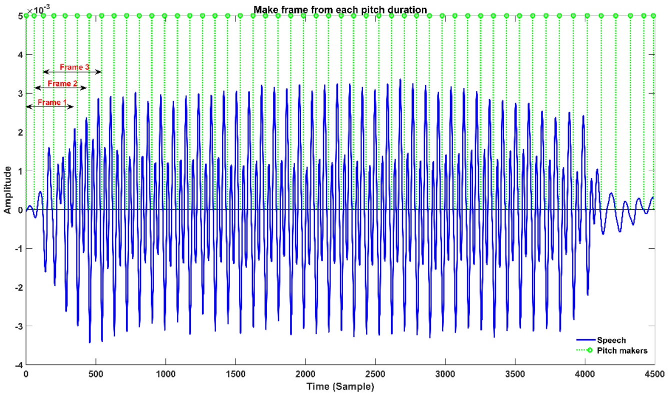


Figure 9. Frame size based on pitch duration of five units, with overlapping of four units.

In the process of creating the frames using a duration of five pitch durations, each frame of the voice signal was multiplied by the Hamming window function. This windowing technique helps to achieve a smoother spectral representation of the signal, reducing the effects of spectral leakage. Figure 10a) depicts the voice signal frames after applying the Hamming window. Once the voice signal frames were obtained, the cepstrum coefficients on the

Mel scale were calculated for each frame. This calculation resulted in a set of cepstrum coefficients on the Mel scale, representing the spectral characteristics of the frame. Figure 10b) illustrates the cepstrum coefficients on the Mel scale for one frame of the voice signal. For each frame of the voice signal, there were M cepstrum coefficients on the Mel scale. As a result, the voice signal produced a total of 44 voice signal frames, as shown in Figure 9. This leads to a dataset of feature values in the frequency domain with dimensions of 44 columns (corresponding to the frames) and M rows (corresponding to the cepstrum coefficients), as shown in Figure 11.

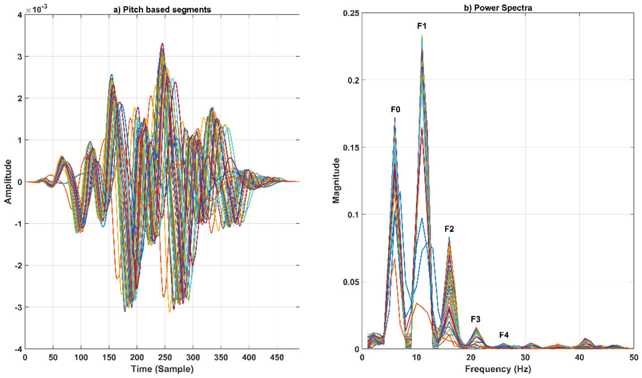


Figure 10. Outcome of frame sizing for the voice signal is 44 frames. a) One frame per five pitch duration. b) Power spectra in each frame signal.

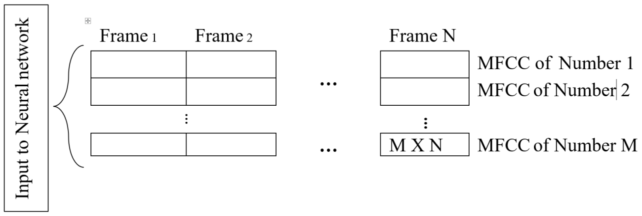


Figure 11. Dataset of feature values in the frequency domain of one voice signal.

As depicted in Figure 11, the data set consists of key attribute values in the frequency domain, representing one sound. In order to process all the sounds, each sound needed to go through the process of identifying features in the frequency domain.

D. Speech Recognition Method

The research employed a multilayer neural network for speech recognition, which is a common approach in this field. The input data for the neural network consisted of two types of parameters: features in the time domain and the cepstral coefficients on the Mel scale obtained from the spectral analysis. The speech recognition performance of a neural network generally depends on a number of factors, such as the structure of the neural network, the datasets used to teach neural networks and input data formats.

1) Determination of the Size of the Input Data: The size of the input data, or the number of recognition parameters, is a crucial factor that can impact the performance of a neural network in speech recognition. It is generally challenging to determine the optimal size of the input data that yields the best recognition accuracy. To find the optimal number of parameters, a gradual increase in the number of parameters is performed, and the recognition efficiency of the neural network is measured. The goal is to identify the number of parameters that results in the highest recognition accuracy. In this research, the number of parameters for the features in the time domain and the size of the cepstral coefficients on the Mel scale were adjusted and used as the input data for the neural network to determine the optimal size of the input data.

2) Input Data Set: The input data set for a neural network consists of the data that is fed into the network for training or recognition testing. In this research, various feature values of speech signals were selected to create input data sets. These feature values were arranged as input vectors to be fed into the neural network. The format of the feature values used to compose the input vectors has a significant impact on the recognition performance of the network. Different combinations of feature values, such as using only parameters of feature values in the time domain, using the cepstral coefficients on the Mel scale, or using both time domain parameters and cepstral coefficients, were employed to create input data sets. The speech recognition performance was then tested using these different input datasets.

3) Architecture of the Neural Network: Since the feature values were extracted from both the time and frequency domains, the formats of the input data inherently differed. Therefore, it was crucial to design a neural network architecture that could effectively accommodate input data with diverse formats. In this research, two different neural network architectures were utilized. The first architecture, referred to as NN1, aimed to evaluate the integration of time and frequency domain features. Its structure is illustrated in Figure 12. The second architecture, known as NN2, was designed to distinguish between feature sets in the time and frequency domains. The structure of NN2 is depicted in Figure 13.

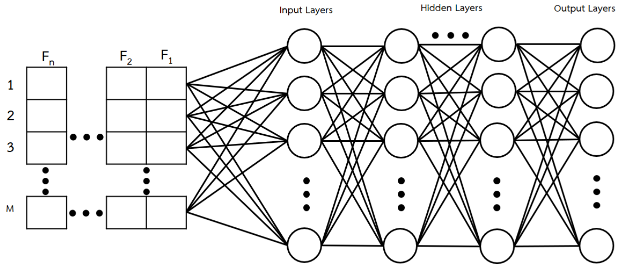


Figure 12. Neural network architecture: NN1 type.

The NN1 neural network architecture consisted of an input layer that matched the number of feature inputs, 300 nodes in the hidden layers, and 50 nodes in the output layer. This architecture was utilized in experiments involving features from both the time domain and the frequency domain. In the context of the architecture, F represented speech signal frames, n represented the sequence of frames, and M denoted the number of key feature parameters in both the time and frequency domains. Figure 12 depicted the NN1 architecture, which allowed for testing the performance of two types of features. The first type involved using only features in the time domain, while the second type combined features from both the time and frequency domains (as illustrated in Figure 13). The transfer function employed in NN1 was as follows: a Sigmoid function was applied to the input layer and hidden layers of each node, while the output layer used a pure linear function.

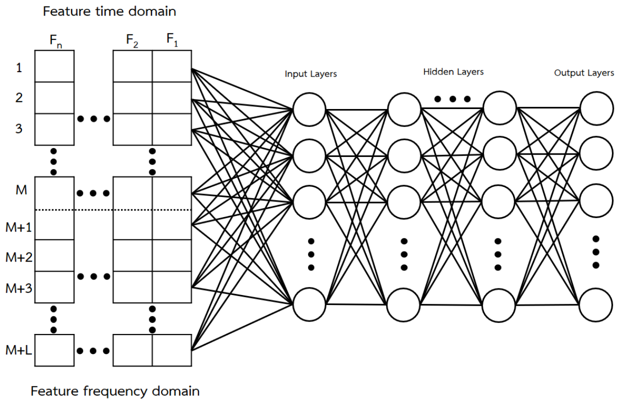


Figure 13. Combination of the feature values sets into one data set.

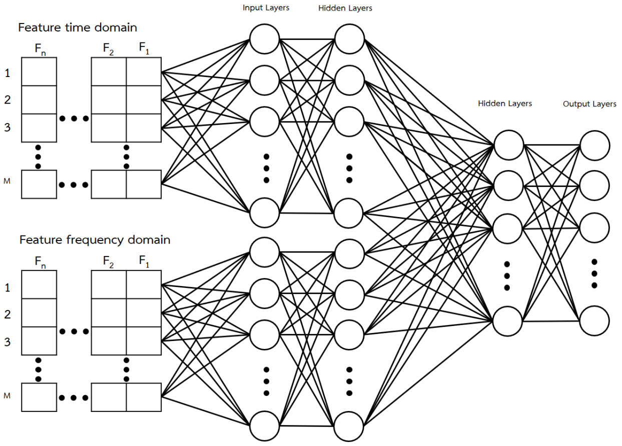


Figure 14. Neural network architecture: NN2 type.

The research explored the combination of feature data sets from the time domain and the frequency axis by concatenating them while using the same NN1 architecture. The assumption was that speech signals undergo continuous changes over time and exhibit periodicity. However, existing methods that utilize the cepstrum coefficient on the Mel scale, obtained through Fourier transform theory, do not account for changes in the signal's period. The objective of the research was to investigate whether incorporating data that captures period changes, along with the cepstrum coefficient on the Mel scale, could improve speech recognition. To test this, the two data sets were merged and employed in speech recognition experiments. In this context, F represented speech signal frames, n denoted the sequence of frames, M represented the number of parameters for feature values in the time domain, and L represented the number of parameters for feature values in the frequency domain (as depicted in Figure 13).

The NN2 architecture consisted of a combination of three neural networks. The concept behind this architecture was as follows: the first neural network received input data containing feature values from the time domain, while the second neural network received input data containing feature values from the frequency domain. The output of the second neural network served as the input for the third neural network, as illustrated in Figure 14

3. Experiments & Results

The speech recognition experiments aimed to assess the performance of neural networks using two sets of feature values. The first set comprised feature values in the time domain, such as position, amplitude values of local minimums and local maximums, and max slope. The second set included feature values in the frequency domain obtained from MFCC, which involved dividing the speech signal into frames based on pitch length. To evaluate the speech recognition performance, three data sets were utilized. The first data set contained feature values from the time domain, the second data set consisted of feature values from the frequency domain, and the third data set combined both types of feature values. These data sets were tested using the NN1 type neural network, depicted in Figure 12 and Figure 13. Furthermore, the second neural network was evaluated using feature data sets from both the time domain and the frequency domain, as shown in Figure 14.

A. *Result of Using Feature Values in the Time Domain with the NN1 Type*

To optimize speech recognition performance, the number of parameters for feature values in the time domain was varied and evaluated. Two parameters, A (signal amplitude) and C (position in the time axis), were considered for each frame. The choice of the number of parameters per frame had an impact on the recognition performance, and it was necessary to identify the optimal number of parameters. In the experiment, the number of feature parameters was adjusted from 4 to 40 points, with an increment of 4 points. Each point was represented by two parameters, A and C. Multiple data sets were utilized for the training process, including sets created solely from male volunteers' speeches, solely from female volunteers' speeches, and from speeches of both genders. The groups of volunteers ranged from 10 people (5 males and 5 females) to 40 people (20 males and 20 females), resulting in four training datasets: 10, 20, 30, and 40 individuals. In each group, the parameters were adjusted as mentioned. The number of parameters that yielded the best results in each group is shown in Table I.

TABLE I. Percentage of Speech Recognition Accuracy of The NN1 Type by Using a and c Parameters of Time Domain Features, along with the Number of Parameters Which Gave the Best Results in Each Training Set (in Bold)

NN1 using parameter A and C features			
training set from speech of male and female volunteer	Number of parameter A and C	Percentage of speech recognition accuracy	
		Traing	Blind
10	32	36.86	11.89
20	32	42.09	22.91
30	32	64.43	43.99
40	**36**	**77.32**	**63.62**

The results presented in Table I show that the speech recognition performance varied based on the parameters A and C used. Among the training datasets, the one with 40 volunteers, including both males and females, and using 36 parameters of A and C, demonstrated the highest accuracy in speech recognition. Specifically, setting the parameters A and C at 32 resulted in the best performance. These findings indicate a positive relationship between the number of volunteers in the training dataset and the overall speech recognition performance, suggesting that increasing the number of volunteers leads to improved results.

B. *Outcome of Utilizing Feature Values in the Frequency Domain with MFCC on the NN1 Type*

In this study, the speech signals were divided into frames based on pitch, allowing for the extraction of MFCC values from these pitch-divided speech signals. These MFCC values were then used as input data for the neural network during the speech recognition process. Given this methodology, it was essential to establish suitable criteria to optimize the performance of speech recognition.

The experiments involved using various data sets for the training process, including separate male speech data sets, female volunteer speech data sets, and combined male and female speech data sets. The groups of volunteers ranged from 10 individuals (5 males and 5 females) to 40 individuals (20 males and 20 females). Additionally, the MFCCs were adjusted and used as input data for the neural network, which had a range of nodes from 4 to 40 (increasing by 4 nodes in each step). This resulted in four groups of training data sets: 10, 20, 30, and 40. The parameters were adjusted accordingly in each group. The optimal number of parameters, yielding the best results, for each group is presented in Table II.

TABLE II. Percentage of Speech Recognition Accuracy of the NN1 Type by Using MFCC Feature in Frequency Domain, along with the Number of MFCC with Best Results in Each Training Set (in Bold)

NN1 using MFCC features			
Training set from speech of male and female volunteers	*Number of MFCC*	*Percentage of speech recognition accuracy*	
		Train	*Blind*
10	20	56.97	44.89
20	16	66.71	56.41
30	16	77.30	67.61
40	**16**	**80.69**	**71.11**

The speech recognition experiments utilizing MFCCs as feature values revealed that the training data set comprising 40 volunteers, including both males and females, and utilizing 16 cepstrum coefficients on the Mel scale, achieved higher accuracy compared to the training data sets with 30 and 20 volunteers. The best performance was observed when utilizing 16 cepstrum coefficients on the Mel scale. These results suggest that increasing the number of volunteers, similar to incorporating feature values in the time domain, leads to improved speech recognition performance.

C. *Results of Applying the A and C Parameters Combined with the MFCC to the NN1 Type*

The results presented in Tables I and II highlight the significant influence of the number of feature values on speech recognition accuracy. The experiments revealed that utilizing 16 cepstrum coefficients on the Mel scale resulted in the highest recognition accuracy, while using 36 feature values in the time domain also achieved the highest accuracy. Building upon these findings, additional testing was conducted to determine the optimal combination of time domain feature values and cepstrum coefficients on the Mel scale for speech recognition. Specifically, 16 cepstrum coefficients were selected, while the number of time domain feature values, represented by parameters A and C, varied from 4 to 36, as presented in Table III.

TABLE III. Recognition Efficacy When Using Parameters A and C of Time Domain Feature Combined with the MFCC as the Speech Recognition Key Attribute Value in NN1 Type

NN1 type		
MFCC+parameter A,C=Feature input size	*Result*	
	Train	*Blind*
16+4=**20**	91.38	85.29
16+8=**24**	91.56	85.34
16+12=**28**	91.71	85.42
16+16=**32**	91.79	85.63
16+20=**36**	92.00	85.71
16+24=**40**	92.21	86.21
16+36=**52**	92.88	83.57

As shown in Table III, the combination of cepstrum coefficients on the Mel scale with parameters A and C yielded a notable improvement in the speech recognition performance of the neural network.

During the testing phase with the training data sets, the highest recognition accuracy of 92.88% was achieved by using 16 cepstrum coefficients on the Mel scale and 36 parameters of A and C as feature values in the time domain. However, in the blind test, the best recognition accuracy of 86.21% was obtained by utilizing 16 cepstrum coefficients on the Mel scale and 24 parameters of A and C as feature values.

It is noteworthy that the combination of 16 cepstrum coefficients on the Mel scale and 36 parameters of A and C did not yield the best results in the blind test, despite its superior performance during testing with the training data set. One possible explanation for this discrepancy is that a higher number of parameters could make the learning process more complex and time-consuming, thereby affecting the recognition accuracy in real-world scenarios.

D. Results of Applying Feature Values in the Time Domain and Frequency Domain to the NN2 Type

The experiment involved training three interconnected neural networks using feature values from both the time domain and the frequency domain. The first neural network, N1, received input data of feature values in the time domain, while the second neural network, N2, received input data of feature values in the frequency domain. The outputs of N1 and N2 were then used as inputs for the third neural network, N3, as shown in Figure 15.

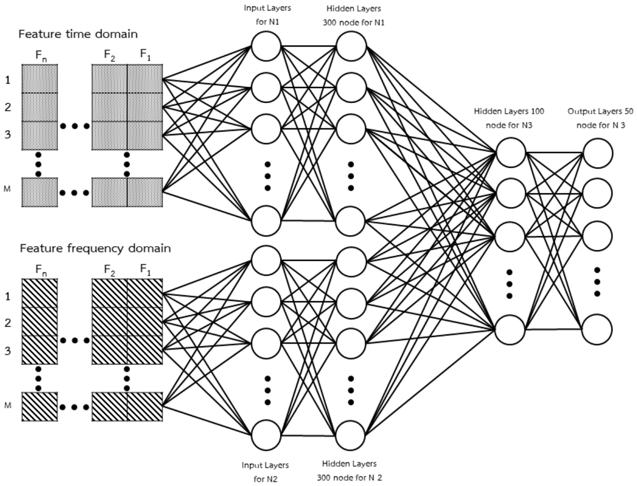


Figure 15. Defining neural network input data from two data sets.

Each neural network had specific configurations: N1 had an input layer matching the size of the time domain feature values, a hidden layer with 300 nodes, and an output layer with 100 nodes. Similarly, N2 had an input layer matching the size of the frequency domain feature values, a hidden layer with 300 nodes, and an output layer with 100 nodes. The outputs of both networks served as the hidden layer for N3, which had 100 nodes, and the final output layer consisted of 50 nodes. Figure 15 depicts the process of feeding the data sets of feature values from the time and frequency domains into individual neural networks.

Due to the distinct formats of feature values from different domains, using a single neural network to effectively incorporate both domains can be challenging. To address this, the proposed approach in this research involved training separate neural networks for each type of feature value. The outputs of these networks were then combined and further trained in another neural network, allowing for the integration of weighted values and adjustment of the output based on the desired number of data groups.

TABLE IV. Speech Recognition Efficacy When Using Parameters A and C of Feature Values in the Time Domain Combined with the Mel Scale Cepstrum Coefficient for Importing to the Neural Network Architecture: NN2

NN2 Type			
feature input size		Output by N3	
number of MFCC input to N2	Number of parameter A,C input to N2	Train	Blind
16	4	93.64	87.30
	8	93.82	87.35
	12	**96.97**	**88.43**
	16	94.05	87.64
	20	94.26	87.72
	24	92.47	87.22
	36	92.14	84.58

The results presented in Table IV highlight the notable enhancement in speech recognition performance by incorporating cepstrum coefficients on the Mel scale and A and C parameters of feature values in the time domain into neural network architecture B. The best recognition outcomes were achieved when utilizing 16 cepstrum coefficients on the Mel scale and 12 parameters of A and C as feature values. The recognition accuracy percentage reached 96.97 for the training data set test and 88.43 for the blind test. These findings demonstrate the effectiveness of this particular combination of feature values in improving speech recognition performance.

4. Conclusion

In this research, the application of a multilayer neural network for speech recognition was explored using different types of feature values, including parameters A and C from the time domain, cepstrum coefficients on the Mel scale obtained through MFCC analysis of pitch-based speech segments, and a combination of MFCC pitch-based segments with parameters A and C. Two neural network architectures, NN1 and NN2, were evaluated. The highest recognition accuracy of 88.43% (blind test) was achieved by NN2, which employed 16 cepstrum coefficients on MFCC pitch-based segments and 12 parameters A and C. NN1, utilizing MFCC pitch-based segments along with parameters A and C, attained the highest recognition accuracy of 86.21% (blind test). Moreover, using only MFCC pitch-based segments or only parameters A and C resulted in recognition accuracies of 71.11% and 63.62% (blind test), respectively. These findings highlight the effectiveness of combining parameters A and C from the time domain with cepstrum coefficients on the Mel scale derived from pitch-based segments in enhancing speech recognition performance.

5. Acknowledgements

This research received financial support from Thailand Science Research and Innovation and Uttaradit Rajabhat University.

References

[1] A. Singh, A. Kittur, K. Sonawane, A. Singh, and S. Upadhya, "Analysis of Time Domain Features of Dysarthria Speech," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 122–125.

[2] R. Jahangir, Y. W. Teh, U. Ishtiaq, G. Mujtaba, and H. F. Nweke, "Automatic Speaker Identification through Robust Time Domain Features and Hierarchical Classification Approach," in *Proceedings of the International Conference on Data Processing and Applications*, ser. ICDPA 2018. New York, NY, USA: Association for Computing Machinery, 2018, pp. 34–38.

[3] M. Sher, N. Ahmad, and M. Sher, "TESPAR feature based isolated word speaker recognition system," in *18th International Conference on Automation and Computing (ICAC)*, Sep. 2012, pp. 1–4.

[4] S. M. Qaisar, S. Laskar, M. Lunglmayr, B. A. Moser, R. Abdulbaqi, and R. Banafia, "An Event-Driven Approach for Time-Domain Recognition of Spoken English Letters," in *2019 5th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP)*, May 2019, pp. 1–4.

[5] "Speaker-independent isolated word recognition using dynamic features of speech spectrum | IEEE Journals & Magazine | IEEE Xplore." [Online]. Available: https://ieeexplore.ieee.org/document/1164788

[6] D. Anggraeni, W. S. M. Sanjaya, M. Y. S. Nurasyidiek, and M. Munawwaroh, "The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm," *IOP Conference Series: Materials Science and Engineering*, vol. 288, no. 1, p. 012042, Jan. 2018, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/288/1/012042

[7] H. Trang, T. H. Loc, and H. B. H. Nam, "Proposed combination of PCA and MFCC feature extraction in speech recognition system," in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, Oct. 2014, pp. 697–702, iSSN: 2162-1039.

[8] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 271–275, Jun. 2015. [Online]. Available: https://doi.org/10.1007/s10772-014-9267-z

[9] F. Z. Chelali and A. Djeradi, "Text dependant speaker recognition using MFCC, LPC and DWT," *International Journal of Speech Technology*, vol. 20, no. 3, pp. 725–740, Sep. 2017. [Online]. Available: https://doi.org/10.1007/s10772-017-9441-1

[10] S. Wiriyarattanakul and N. Eua-anant, "Accuracy Improvement of MFCC Based Speech Recognition by Preventing DFT Leakage Using Pitch Segmentation," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 1-8, pp. 173–179, Feb. 2018, number: 1-8. [Online]. Available: https://jtec.utem.edu.my/jtec/article/view/3756

[11] J.-w. Zhu, S.-f. Sun, X.-l. Liu, and B.-j. Lei, "Pitch in Speaker Recognition," in *2009 Ninth International Conference on Hybrid Intelligent Systems*, vol. 1, Aug. 2009, pp. 33–36.

[12] M. B. Jdira, I. Jemâa, and K. Ouni, "Speaker recognition system based on pitch estimation," in *2014 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, 2014, pp. 1–5.

[13] A. A. Khulage, "Extraction of pitch, duration and formant frequencies for emotion recognition system," in *Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)*, Oct. 2012, pp. 7–9.

[14] M. kammoun and N. Ellouze, "Pitch and Energy Contribution in Emotion and Speaking styles Recognition Enhancement," in *The Proceedings of the Multiconference on "Computational Engineering in Systems Applications"*, vol. 1, Oct. 2006, pp. 97–100.

[15] X. Xu, T.-q. Zhang, S. Shi, and Y.-j. Zhang, "An improved pitch detection of speech combined with speech enhancement," in *2014 7th International Congress on Image and Signal Processing*, Oct. 2014, pp. 778–782.

[16] Q. Huang, D. Wang, and Y. Lu, "Single channel speech enhancement based on prominent pitch estimation," in *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2009)*, Dec. 2009, pp. 205–208.

[17] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Speech enhancement by harmonic modeling via map pitch tracking," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I–549–I–552, iSSN: 1520-6149.

[18] D. A. Eddins, S. Anand, A. Camacho, and R. Shrivastav, "Modeling of Breathy Voice Quality Using Pitch-strength Estimates," *Journal of Voice*, vol. 30, no. 6, pp. 774.e1–774.e7, Nov. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199715002672

[19] S. McLaughlin, D. Leith, and I. Mann, "Using Gaussian processes to synthesise voiced speech with natural pitch variations," in *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)*, vol. 1, Jul. 2002, pp. 321–324 vol.1.

[20] L. Ru-Wei, C. Long-Tao, and L. Yang, "Pitch Detection Method for Noisy Speech Signals Based on Wavelet Transform and Autocorrelation Function," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Oct. 2013, pp. 153–156.

[21] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002. [Online]. Available: https://doi.org/10.1121/1.1458024

[22] S. Wiriyarattanakul and N. Eua-anant, "Pitch segmentation of speech signals based on short-time energy waveform," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 907–917, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s10772-017-9459-4

Dr.Sopon Wiriyarattanakul He received the Bachelor of Engineering Program in Computer Engineering from Naresuan University, Phitsanulok, Thailand, in 2005. The Master of Engineering in Computer engineering from Chiang Mai University, Chiang Mai, Thailand, in 2009 and the Doctor of Philosophy Program in Computer Engineering from Khon Kaen University, Khon Kaen, Thailand, in 2018. His research interest includes data analysis, machine learning, computer vision, , natural language processing, signal processing, and pattern recognition.

Dr.Piroon Kaewfoongrungsi He received the Bachelor of Science in Technical Education (Electronics and Computer) from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 2004. The Master of Engineering in Computer engineering from Chiang Mai University, Chiang Mai, Thailand, in 2009 and the Doctor of Philosophy Program in Computer Engineering from Khon Kaen University, Khon Kaen, Thailand, in 2018. His research interest includes internet of thing, machine learning, microcontroller, ECG signal processing, and Robot.

Asst.prof. Ekkalak Sumonphan He received the Bachelor Degree of Computer Engineering from Institute of Rajamangala University of Technology, Phathumtani, Thailand, in 2002 and the Master of Engineering in Computer engineering from Chiang Mai University, Chiang Mai, Thailand, in 2009. His research interest includes internet of thing, machine learning, Computer Vision, and Embedded System.