Audio Feature and Correlation Function-Based Speech Recognition in FM Radio Broadcasting

Narathep Phruksahiran[†], Non-member

ABSTRACT

The analysis and classification of audio signals are becoming increasingly important, especially in the age of communication and dissemination of information through radio broadcasting systems. It is therefore essential that systems and platforms are available to monitor the spread of fake or fraudulent news. A speech feature-based correlation (SFC) algorithm and a speech recognition framework are developed in this study, combining specific speech features and performance correlation to monitor real-time radio broadcasting and recognize specific speech based on human samples. The speech features include the Mel frequency cepstral coefficient, gammatone cepstral coefficient, spectral entropy, and pitch. The results illustrate the advantages and disadvantages of each feature applied to the various speech sound groups. Furthermore, each feature combined with the design of SFC further enhances system performance and increases accuracy.

Keywords: Speech Recognition, FM Radio Broadcasting, Audio Feature, Correlation Function

1. INTRODUCTION

Current technological developments are likely to be centered around the synergy between machines and humans through various control methodologies. For security reasons, commands require a password or fingerprint authorization to manage interactions. However, another unique, widely used feature is voice control, or a technology called machine hearing, which aims to sense an acoustic environment in the same way as humans do. Sound signals can commonly be divided into music, speech, and environmental acoustics, depending on the instrument and language type.

Speech data are an essential communication tool for connecting humans and allowing them to exchange their knowledge and viewpoints. One of the fundamental technologies is a radio network for broadcasting news. Currently, radio stations are a practical means of spreading valuable information, such as news and entertainment. Nevertheless, some radio stations broadcast news to society in a negative way. Government agencies need to use modern technologies to monitor radio stations, including analyzing the information broadcasted. These systems must be flexible, fast, and effective.

1.1 Related Work

The analysis of sound signals, also known as automatic sound recognition (ARS), is being continuously developed and can be divided into research groups such as speech detection, speaker recognition, and speech recognition [1–8]. Signal processing, feature extraction, and classification are at the core of the ARS system. This essential information can be used to analyze audio signals and music classifications [9, 10]. Research on speech recognition focuses on simplifying operations and making them more flexible through correlation techniques without applying complex learning algorithms [11–13].

In recent times, online media has been used to spread fake news and hate speech, and many researchers aim to detect and monitor user posts before they spread [14–16]. Profanity and offensive speech used in online media have literal characteristics that can be used to find prototypes and comparisons. However, broadcasting via a radio station is more complex. In addition, the radius of radio propagation is limited depending on the transmission power of each radio station. It is therefore necessary to design and develop a system for monitoring radio stations [17–19].

The voice of each country's language is unique. Some languages have the same tone of voice, while others have a tone of speech. From a study on the unique characteristics of the Thai language, the Thai language is identified as having five distinctive tones, each represented by a single fundamental frequency, as demonstrated by [20–22]. It is necessary to understand the nature of each region's sound characteristics because speech analysis is essentially interpreted in both the time and frequency domain.

1.2 Motivation and Contributions

The measurement platform and speech recognition method developed in this study are designed to detect speech with specific messages sent through FM radio propagation. The method has been developed for security purposes and to create a real-time broadcasting monitoring system to address the current engineering challenges. The developed speech feature-based correlation (SFC)

Manuscript received on November 10, 2021; revised on March 17, 2022; accepted on June 7, 2022. This paper was recommended by Associate Editor Piya Kovintavewat.

The author is with the Department of Electrical Engineering, Chulachomklao Royal Military Academy, Nakhon Nayok, Thailand.

[†]Corresponding author: narathepp@gmail.com

Digital Object Identifier: 10.37936/ecti-eec.2022203.247516

algorithm consists of two processing steps, the first of which requires the relevant thresholds to be determined based on the correlation coefficients of sample speech. The second step involves comparing the correlation coefficients of selected features both in the frequency and time domain.

This paper proposes a speech recognition framework and demonstrates its performance using the speech detection application in real-time FM radio broadcasts. The main contributions of this paper are summarized as follows. (1) In contrast to the existing methodology, this paper proposes a recognition framework that combines the classification performance of the correlation function with specific speech features to detect particular word sentences in real-time FM radio broadcasts. (2) Under the SFC method, specific speech features are proposed, such as the Mel frequency cepstral coefficient (MFCC), gammatone cepstral coefficient (GTCC), spectral entropy (SE), and pitch (P). (3) Extended operations with real-time FM radio broadcasting are conducted in this paper. The outcomes confirm the efficiency of the proposed algorithm in terms of its detection performance and scalability.

This paper is organized as follows. In Section 2, the system model for FM signals, speech feature extraction, correlation, and hardware implementation are briefly introduced as the basic theory of the SFC algorithm. Section 3 describes the SFC algorithm framework. Section 4 discusses the experimental results. Finally, Section 5 presents a summary of the paper and the conclusions reached.

2. SYSTEM MODEL

2.1 FM Signal

Frequency modulation is the angle modulation of a carrier oscillation according to the changes in the modulating signal. With frequency modulation, the amplitude of the carrier frequency remains unaffected. The instantaneous value of the information signal (NF) amplitude corresponds to the respective volume in sound transmissions. The frequency change is called the frequency deviation. The greater the amplitude of the modulating NF signal is, the greater the frequency deviation. According to [23], the FM signal equation can be written as

$$u_{FM}(t) = A\cos(\varphi(t)) = A\cos\left(\varphi\left[u_{NF}(t)\right]\right) \qquad (1)$$

where $u_{NF}(t) = A\sin\left(\omega_{NF}t\right)$ is the data to be modulated. The derivative of the phase angle of an angle-modulated oscillation is called the instantaneous frequency and is written as

$$\omega(t) = \frac{d\varphi(t)}{dt}. \qquad (2)$$

When $\omega = 2\pi f$, the following instantaneous frequency is obtained:

$$f(t) = \frac{1}{2\pi}\frac{d(\varphi(t))}{dt}. \qquad (3)$$

According to the frequency modulation, a linear relationship exists between the phase and the signal's integral; written as

$$\varphi_{FM}(t) = \omega_T t + K_{FM}\int_0^T u_{NF}(t)\,dt. \qquad (4)$$

The modulation constant K_{FM} has the unit of second per volt. The ratio of the frequency deviation Δf to the signal frequency f_{NF} is called modulation index η. For simplification, the complex voltage of the modulated signal $u_{FM}(t)$ is written as

$$u_{FM}(t) = e^{-j\left(\eta\cos(\omega_{NF}t-\omega_T t\right)} \qquad (5)$$

From the complex representation, the following equation is obtained:

$$\alpha(t) = \arctan\frac{\mathrm{Im}\left(\mathrm{u_{FM}}(\mathrm{t})\right)}{\mathrm{Re}\left(\mathrm{u_{FM}}(\mathrm{t})\right)} \qquad (6)$$

The demodulated signal $u_{NF}(t)$ is obtained by deriving the instantaneous phase as

$$u_{NF}(t) = k_{Dem}\frac{d}{dt}\alpha(t) \qquad (7)$$

where k_{Dem} refers to the demodulation constant. In this research, the detected FM radio signals undergo a demodulation process using the comm.FMBroadcastDemodulator module in MATLAB. The developed program demodulates a complex baseband FM signal and filters it to create an audio signal. The expanded program has also been modified to the carrier frequency and bandwidth to conform with the radio station used in the experiments.

2.2 Speech Feature Extraction

The four existing acoustic features utilized in this paper are briefly presented in this section, further details of which can be found in the references.

2.2.1 Mel Frequency Cepstral Coefficient (MFCC)

The MFCC is the most commonly used feature in automatic speech recognition, as presented in [27]. The MFC feature extraction procedure incorporates windowing the signal to produce a framed signal, implementing the DFT to transform the signal into the frequency domain, and then mapping the frequencies on the Mel Scale. The cepstral parameters are then calculated from the filter bank amplitude logarithm by implementing the inverse discrete cosine transform (DCT). Finally, the MFCC can be formulated as

$$MFCC_m = \sum_{n=1}^{N} X_n\cos\left[\frac{\pi m}{N}\left(n-\frac{1}{2}\right)\right] \qquad (8)$$

where $1 \le m \le M$, X_n represents the log-energy output of the n-th filter, N is the number of cepstrum coefficients, and M is the number of MFCCs.

2.2.2 Gammatone Cepstral Coefficient (GTCC)

The estimation method for gammatone cepstral coefficients is analogous to the MFCC extraction scheme, as demonstrated in [24, 25]. In the first step, the audio signal is windowed into short blocks. After fast Fourier transform (FFT), the signal is applied to the gammatone filter bank, which is the outcome of a gamma distribution and a sinusoidal tone of the center frequency (f_c). The log function and discrete cosine transform (DCT) are then implemented to create a human loudness perception feature, decorrelating the logarithmic-compressed filter output as

$$GTCC_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \log\left(X_n\right) \cos\left[\frac{\pi n}{N}\left(m - \frac{1}{2}\right)\right] \quad (9)$$

where $1 \leq m \leq M$, X_n is the energy of the signal in the n-th spectral band, N is the number of gammatone filters, and M is the number of GTCCs.

2.2.3 Spectral Entropy (SE)

Spectral entropy can be used to estimate the number of bits expected to outline some of the information. As described in [29, 30], when employed in the probability mass function, entropy can also be used to estimate the peakiness of a distribution. The calculation involves dividing the individual frequency components of a spectrum by the sum of its parts.

$$x_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \quad (10)$$

where X_i is the energy of the spectrum's frequency component, $x_i = x_1, \ldots, x_N$ is the probability mass function of the spectrum, and N is the number of spectrum points. For a particular block, the entropy is calculated from x_i by

$$H = -\sum_{i=1}^{N} x_i \log_2 x_i \quad (11)$$

2.2.4 Pitch (P)

The pitch or fundamental frequency is an imperative element of numerous speech processing applications, and various procedures have been described in [31, 32]. As demonstrated by [33], the groups are divided according to whether they are conducted in the time, frequency, or time-frequency domain. The power spectral density at frequency f_0 and time t can be formulated as

$$Y_t(f) = \sum_{k=1}^{K} a_{k,t} \delta\left(f - kf_0\right) + N_t(f) \quad (12)$$

where $N_t(f)$ depicts the power spectral density of the undesired noise and $a_{k,t}$ denotes the k-th harmonic power. In the log-frequency domain, the signal's energy can be determined using the convolution of the signal model and an impulse response of the filter $Y_t(q) * h(q)$ that peaks at $q_0 = \log f_0$ with

$$Y_t(q) = \sum_{k=1}^{K} a_{k,t} \delta\left(q - \log k - \log f_0\right) + N_t(q) \quad (13)$$

$$h(q) = \sum_{k=1}^{K} \delta\left(q - \log k\right) \quad (14)$$

Consequently, the pitch can be obtained by selecting the most prominent peak in the filter's output.

2.3 Correlation

The energy of the differential signal to be compared is used as a measure of the similarity or correlation. According to [34], the signal energy of the actual signal $s(t)$ in the time interval $\left[t_1, t_2\right]$ can be written as

$$E = \int_{t_1}^{t_2} s^2(t)\, dt. \quad (15)$$

Two real energy signals are considered, namely $s(t)$ and $g(t)$, and a measure needs to be identified that describes the similarity or correlation of the two signals. The energy E_Δ of the differential signal $\Delta(t) = s(t) - g(t)$ can be used in this case, written as

$$E_\Delta = \int_{-\infty}^{\infty} [s(t) - g(t)]^2\, dt. \quad (16)$$

To make E_Δ independent from the amplitudes of the signals $s(t)$ and $g(t)$, they are normalized with the help of their energy E_s and E_g, with $s_n(t) = s(t)/\sqrt{E_s}$ and $g_n(t) = g(t)/\sqrt{E_g}$, thereby obtaining

$$E_{\Delta,n} = \int_{-\infty}^{\infty} \left[s_n(t) - g_n(t)\right]^2 dt \quad (17)$$

$$= 2 - 2\frac{\int_{-\infty}^{\infty} s(t)\, g(t)\, dt}{\sqrt{E_s E_g}} \quad (18)$$

The standardized cross-correlation coefficient is defined as the measure of similarity as

$$\rho_{sg} = 1 - \frac{E_{\Delta,n}}{2} = \frac{\int_{-\infty}^{\infty} s(t)\, g(t)\, dt}{\sqrt{E_s E_g}} \quad (19)$$

The normalized value of the cross-correlation coefficient lies in the interval $-1 \leq \rho_{sg} \leq 1$ and becomes 1 for $g(t) = ks(t)$ with a real, positive k.

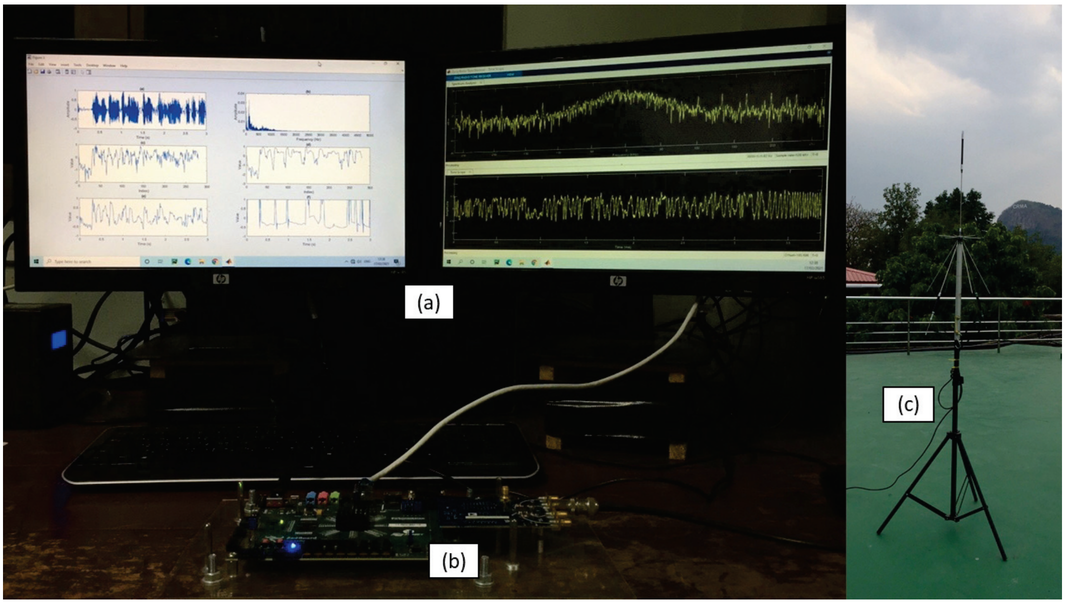

***Fig. 1**: Photograph of the testing environment; (a) control PC, (b) FMCOMMS3 and ZedBoard, and (c) antenna.*

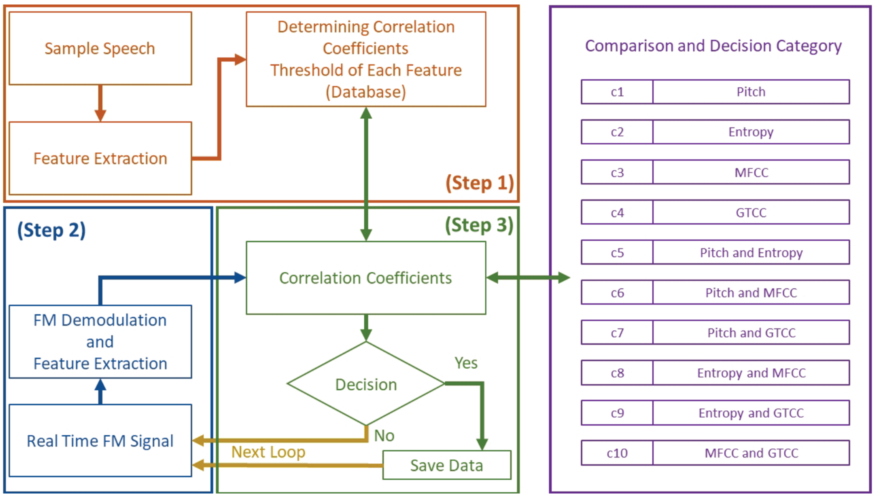

***Fig. 2**: Flowchart of the FM broadcast monitoring system and SFC algorithm.*

***Table 1**: Hardware specifications.*

Parameter	Value
RF transceiver	2×Tx and 2×Rx
Frequency range	70 MHz to 6.0 GHz
Channel bandwidth	<200 kHz to 56 MHz
RF inputs (peak power)	2.5 dBm
Operating temperature	−40 to +85 ℃

2.4 Hardware Implementation

The performance of the proposed SFC algorithm is validated by employing a combination of the Avnet ZedBoard with the AD-FMCOMMS3-EBZ FMC analog device module. Table 1 displays the hardware specifications in the specified range of RF spectra using MATLAB R2019 to implement the proposed algorithms with a 64-bit computer consisting of a Core i5 processor and 4 GB RAM.

Fig. 1 presents the experimental setup, where FMCOMMS3 and the ZedBoard interface are employed in the system through MATLAB software. The receiving antenna is located at 14.303263° N, 101.164968° E, approximately 15 meters above the ground. The AOR DAG735G antenna is connected to the Rx port of the FMCOMMS3 board and can satisfy a frequency scale of 75 MHz to 3 GHz.

3. PROPOSED FRAMEWORK AND ALGORITHM

A flowchart of the FM broadcast monitoring system and SFC algorithm is displayed in Fig. 2. The three main processing steps consist of the developed processing framework and speech feature-based correlation (SFC)

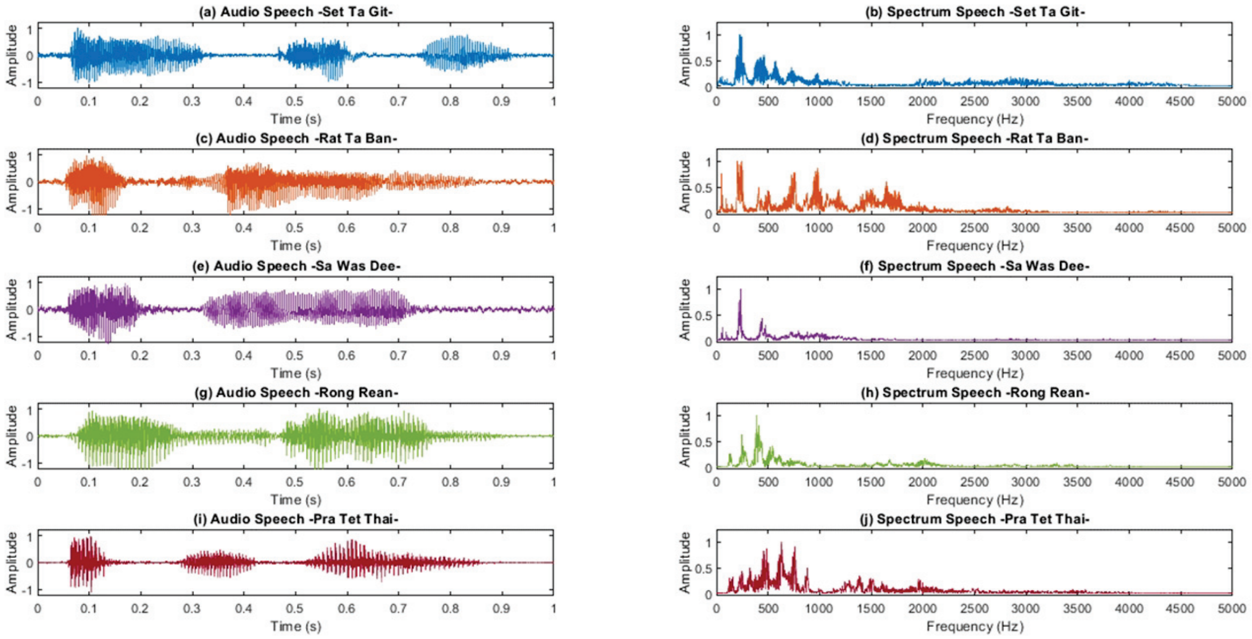

***Fig. 3**: Audio signals of sample speech and the spectrum after the pre-emphasis step.*

algorithm. Step 1 speech feature extraction; Step 2 FM demodulation and feature extraction; and Step 3 correlation and decision-making.

3.1 Step 1: Speech Feature Extraction

The aim of the initial step is to discover a suitable threshold based on the sample speech correlation coefficients as shown in Fig. 2 (Step 1).

In the first stage, audio samples of specific speech used in the speech detection prototype for radio broadcasts are collected and pre-emphasized by a low-pass filter to reduce noise in the speech signal. Five speech messages were selected in this paper, namely, “Set Ta Git” (economic), “Rat Ta Ban” (government), “Sa Was Dee” (hello), “Rong Rean” (school), and “Pra Tet Thai” (Thailand) with 104 samples of each message, spoken in the Thai language, recorded using 15 male and 20 female Thai native speakers.

Fig. 3 depicts an example of each recorded speech in the time domain with a duration of one second for “Set Ta Git,” “Rat Ta Ban,” “Sa Was Dee,” “Rong Rean,” and “Pra Tet Thai,” as shown in Figs. 3(a), 3(c), 3(e), 3(g), and 3(i), respectively. Furthermore, the spectrum of each speech sample can be observed from the time domain data displayed in Figs. 3(b), 3(d), 3(f), 3(h), and 3(i), respectively. The graph characteristics demonstrate the distinction of a signal envelope in the time domain, including the frequency spectrum of each speech signal, which is an essential basis for feature extraction. Pitch (P), Spectral Entropy (SE), Gammatone Cepstral Coefficient (GTCC), and the Mel Frequency Cepstral Coefficient (MFCC) have been chosen in this paper, as described in Section 2.

The rationale for the five terms under study comes from the hypothesis that a radio station might use these words in general conversation. This practical application adapts and examines specific words or messages that violate the rights of others for further analysis. These passages are characterized by the feature extraction process and then compared with the features within each speech group. Fig. 4 illustrates the extracted features derived from the speech sample “Set Ta Git.” Fig. 4(a) presents the audio signal in the time domain. The corresponding spectral, MFCC, GTCC, entropy, and pitch are displayed in Figs. 4(b)–4(f), respectively.

The determined correlation coefficient and threshold of each feature are stored in the database. The resulting correlation coefficients are used as a reference or threshold for examining the real-time signal monitoring of the FM radio station.

3.2 Step 2: FM Demodulation and Feature Extraction

Once the appropriate reference values in Step 1 have been obtained, in Step 2, the system will monitor the FM radio station. This paper uses the FM broadcasting signal from the CRMA FM radio 89.75 MHz, located at 14.279154° N and 101.163686° E, about 3 km away from the proposed system. First, the FMC module captures the real-time FM signal with a one-second duration in each loop, as shown in Fig. 2 (Step 2). The signal then goes through the demodulation process using the program developed in this study. The result is an audio signal with a length of one second. The detected signal is then

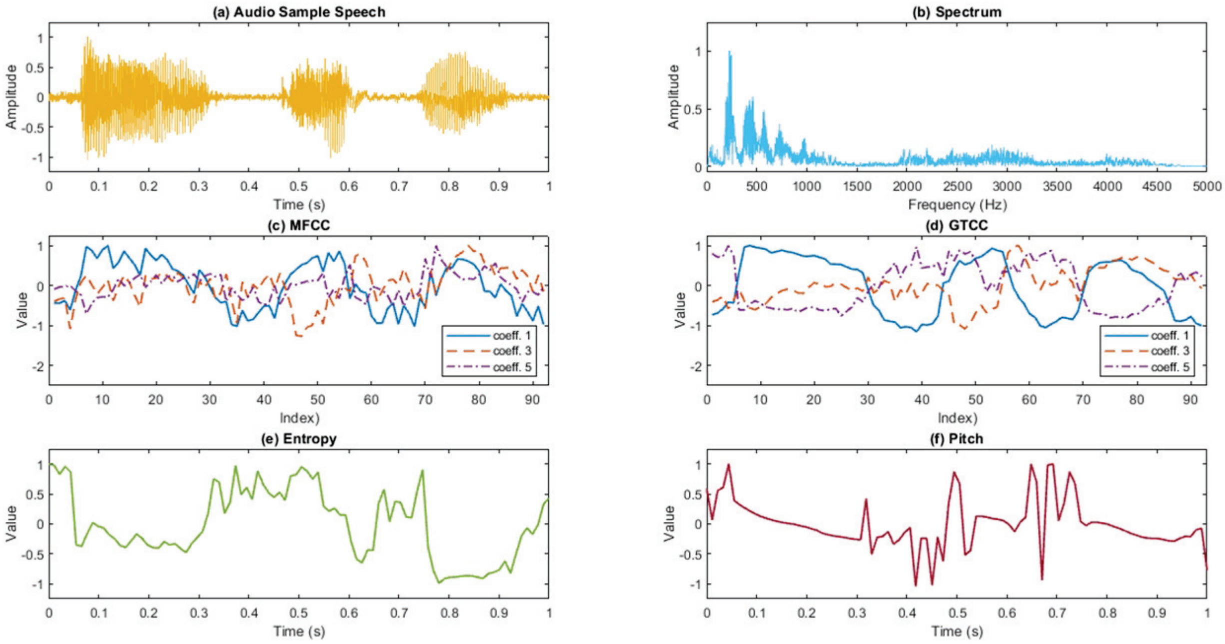

***Fig. 4**: The SFC algorithm modeling sample speech "Set Ta Git" in time and frequency domains; (a) voice speech in the time domain, (b) voice speech in the frequency domain, (c) MFCC, (d) GTCC, (e) spectral entropy, and (f) pitch.*

analyzed to determine the specific features. It should be noted that these are broadcast signals. Thus, there is a general use of words and sounds, not only in the five speech samples. The aim of this paper is to develop a monitoring system to search for specific words under general messages actually broadcast.

3.3 Step 3: Correlation and Decision

The final step includes a comparison of the correlation coefficients to facilitate decision-making, as shown in Fig. 2 (Step 3). Finally, the speech features obtained from the real-time FM broadcast data are compared with the sample speech sets in the database developed for this study.

In this paper, the comparison conditions are classified into ten groups, as shown in Fig. 2 (Comparison and Decision Category). In c1, c2, c3, and c4, the comparison is made using one feature of pitch, entropy, MFCC, and GTCC, respectively. In the case of c5 to c10, combinations of various features are employed. For example, in the case of c5, the requirement is only valid if the pitch and entropy values are sufficient to show that the detected data are similar to the established criteria. This developed process improves the efficiency in verifying more accurate speech recognition.

If the correlation coefficient value is lower than the defined threshold value, the system will skip the data and move to the next loop to monitor the following demodulated FM data. If the value of the variable value is greater than or equal to the predetermined reference value, all the datasets are recorded, and the system will manage the next loop.

4. EXPERIMENTAL RESULTS

The experimental results were divided into two main parts, the first of which involved an experiment to determine the SFP algorithm's effectiveness with five speech samples recorded from humans aged between 20 and 65 years old in a room environment without special audio equipment. The second part uses the developed platform and SFC algorithm to monitor an FM radio broadcast signal and recognize the specifically selected speech.

4.1 Sample Speech

The dataset with five types of original Thai voice signals: "Set Ta Git" (economic, E), "Rat Ta Ban" (government, G), "Sa Was Dee" (hello, H), "Rong Rean" (school, S), and "Pra Tet Thai" (Thailand, T), is analyzed in this section. Each speech sample contains 104 datasets from 15 male and 20 female Thai native speakers. The objective of this experiment was to test the correlations among speech sound datasets and identify appropriate threshold values for each feature to determine the correlation with the radio broadcast datasets. We used speech audio data to analyze the conditions of four specific features: MFCC, GTCC, entropy, and pitch. All the sample speech datasets were used to determine the correlations within the same speech collection with the mean and standard deviation investigated for comparison.

Table 2 compares the mean and standard deviations of each feature's correlation coefficients in each audio set

***Table 2**: Mean and standard deviation of feature correlation coefficients.*

Speech	$\text{MFCC}_{1,\text{m}}$	$\text{MFCC}_{1,\text{s}}$	$\text{GTCC}_{1,\text{m}}$	$\text{GTCC}_{1,\text{s}}$	$\text{Entropy}_{\text{m}}$	$\text{Entropy}_{\text{s}}$	Pitch_{m}	Pitch_{s}
"Set Ta Git" (E)	0.5262	0.1819	0.4172	0.1771	0.3424	0.1421	0.2590	0.1231
"Rat Ta Ban" (G)	0.7282	0.2049	0.4666	0.1539	0.3118	0.1534	0.4014	0.1693
"Sa Was Dee" (H)	0.5451	0.1815	0.3215	0.1550	0.3118	0.1534	0.2467	0.1512
"Rong Rean" (S)	0.4572	0.2054	0.2972	0.1575	0.1689	0.1534	0.3244	0.1091
"Pra Tet Thai" (T)	0.4974	0.1813	0.3669	0.1274	0.2591	0.1690	0.2297	0.1081

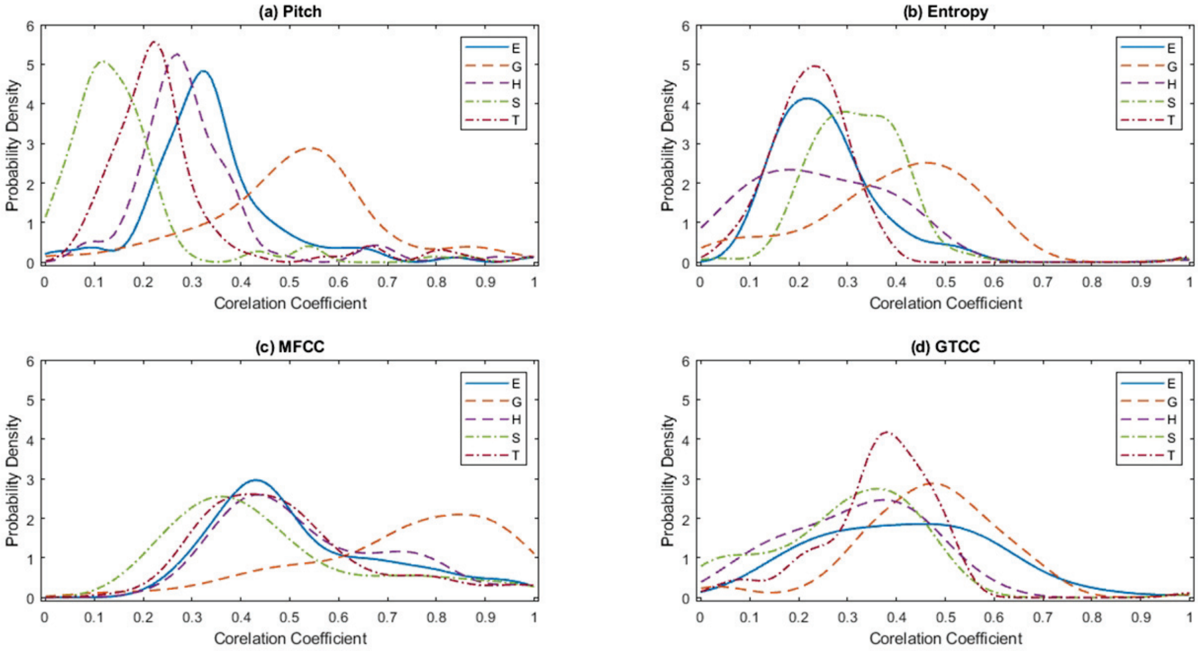

***Fig. 5**: Distribution of the correlation coefficient values; (a) pitch, (b) entropy, (c) MFCC_1, and (d) GTCC_1.*

group. The mean of the correlation coefficients in each speech group is different. For example, in speech group E, the most correlated feature in the dataset was MFCC_1 at 0.5262, while the least correlated feature was pitch at 0.2590. In the G group, the correlation coefficients of the dataset samples were highest when the MFCC_1 feature value of 0.7282 was used and the lowest at an entropy value of 0.3118.

It can be observed that when analyzing the speech signal in combination with various features among the sample set, the results obtained from the MFCC_1 feature exhibit the highest correlation with the dataset, followed by GTCC_1, entropy, and pitch feature. However, they are still considered to have comparable values and can be effectively implemented according to the proposed SFC algorithm.

The distribution of the correlation coefficient values can be expressed as a histogram or probability density function (PDF). For ease of understanding, we fit a probability distribution object to the histogram data of the correlation coefficient values using a kernel distribution. Figs. 5(a)–5(d) show the probability of a particular correlation coefficient in pitch, entropy, MFCC_1, and GTCC_1 speech group, respectively.

It can be observed from the graph that each group's distribution is different in nature and unique to each speech group and language. The distribution shown in Fig. 5 corresponds to the mean and standard deviation shown in Table 2. The diagrams are the basis for determining the threshold value in the development step and display the proportion of speech data that may be mistakenly recognized when the threshold value used for each speech signal group is too high or low. This value is the key to determining the proper threshold value in a real-time application to suit the nature and tone of the language used.

In the next step, the SFC algorithm is applied to separate the speech audio signals into specific classification groups, as presented in Fig. 2. The threshold for the correlation coefficient values was set at the mean to classify the datasets into individual groups (c1–c10).

In Fig. 6, the x-axis represents the classification group while the y-axis represents the percentage of each speech. The factors used to classify the datasets into

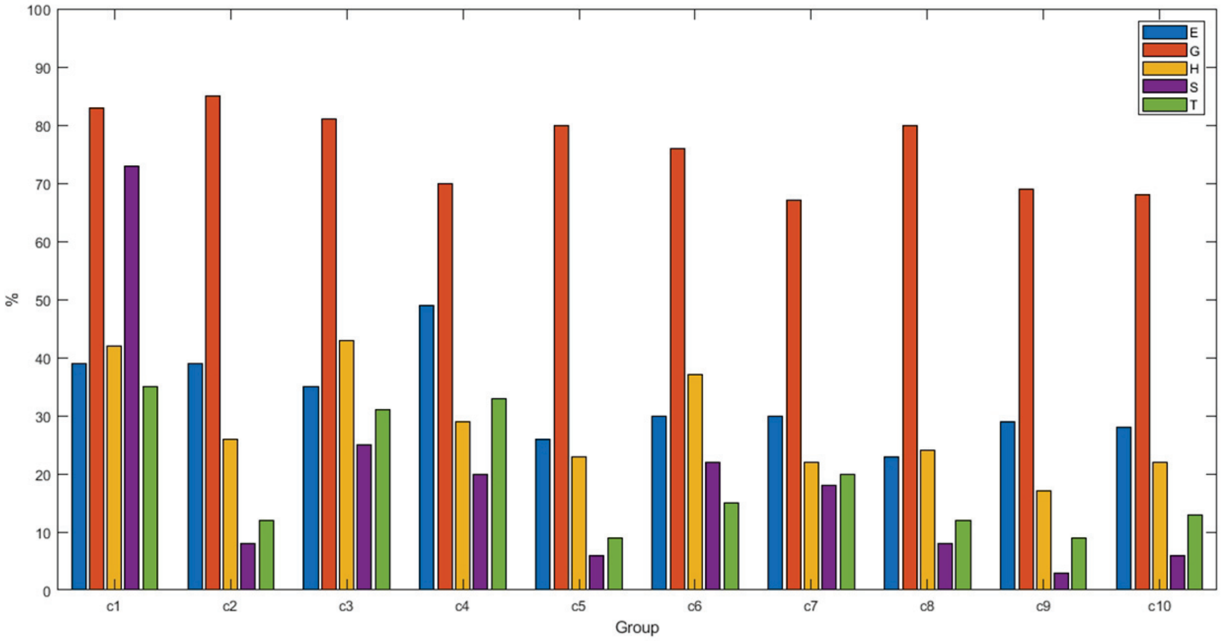

***Fig. 6**: A comparison of the grouping results for each sample speech sound with reference to the mean of the correlation coefficient.*

groups c1, c2, c3, and c4 included only one feature type. If the speech dataset has a correlation coefficient value of greater than or equal to the threshold value, it will be organized into that group. As can be observed from c1, the G speech group has the highest accuracy rate. This indicates that the accuracy of the G sound signal is approximately 80 percent. In addition, the graph indicates that the G sound group has a correlation in pronunciation and speech that can be parsed more appropriately and accurately than the S sound group with the mean as the threshold.

As can be observed, in a group with a small accuracy ratio, such as c2 of the S sound group, the characteristics of the entropy of this speech group are very diverse; hence, applying only the entropy analysis leads to significant errors in real-time implementation. Therefore, it should be complemented by MFCC and GTCC to provide greater accuracy. The results shown in Fig. 6 illustrate the advantages and disadvantages of each feature applied to each sound group. Furthermore, the coordination design of each feature further enhances the system's performance and provides more accurate results. This knowledge is helpful in detecting and increasing accuracy when recognizing specific speech or sentences.

4.2 FM Radio Broadcast

This section presents the experimental results with the developed platforms and algorithms for monitoring the audio signals emitted by FM radio stations. The signals were examined by recording a one-second FM signal frame to process in each cycle according to the SFC algorithm by continuously processing tens of thousands of cycles and collecting statistical data to recognize a particular speech group signal. The speech groups were divided into five groups:

Fig. 7 displays an example of the captured demodulated FM signal, classified as "Set Ta Git." Fig. 7(a) presents the audio signal in the time domain. The related spectral, MFCC, GTCC, entropy, and pitch are demonstrated in Figs. 7(b)–7(f), respectively.

Fig. 8 shows the clustering results of the detectable signals classified by speech groups based on the mean correlation values and grouped by feature matching based on the SFC algorithm. It is clear from Fig. 8 that the system can detect many speech samples, which is consistent with the fact that the station host greets the listeners.

In the experiments with accurate signals, there is a limit to the inability to control the content and information of the radio stations. However, the experiments follow the actual application conditions for which this platform is developed. The system demonstrates the efficiency of monitoring and recognizing specific speech samples or sentences.

5. CONCLUSION

In this paper, a processing framework and speech feature-based correlation (SFC) algorithm are proposed. Five of the most common speech signals are compiled using 15 male and 20 female volunteers. Each audio speech set contains 104 samples. First, the essential

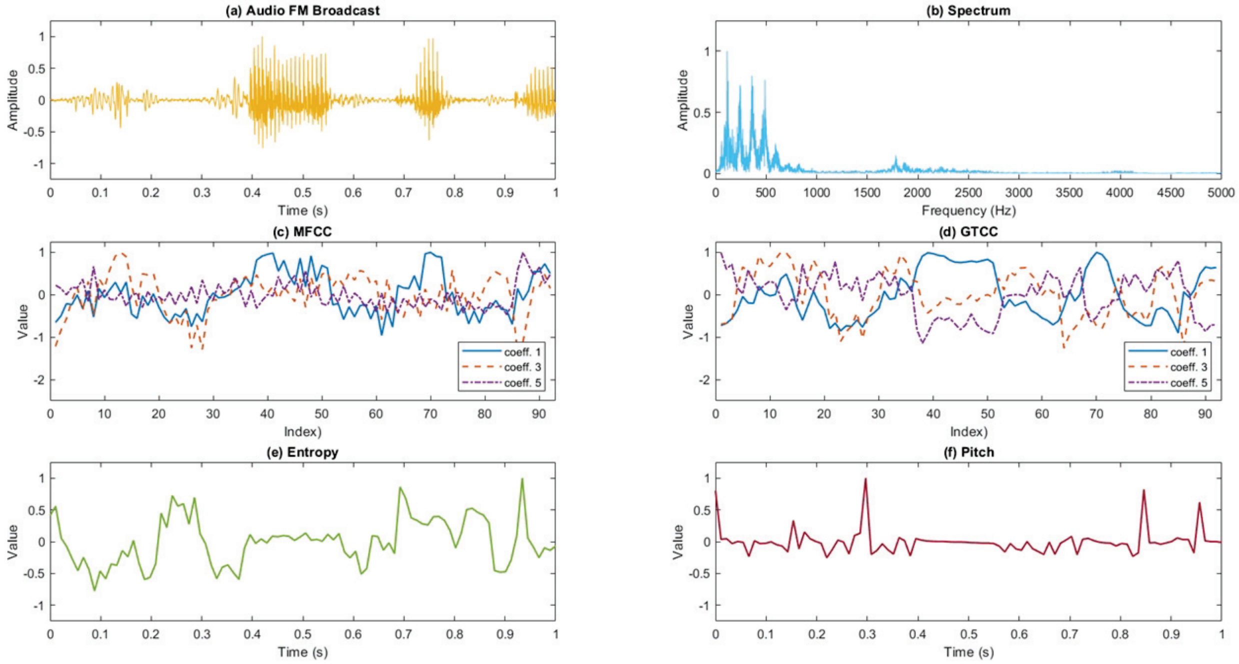

***Fig. 7**: Example of the captured signal, classified as "Set Ta Git" in the time and frequency domains; (a) voice speech in the time domain, (b) voice speech in the frequency domain, (c) MFCC, (d) GTCC, (e) spectral entropy, and (f) pitch.*

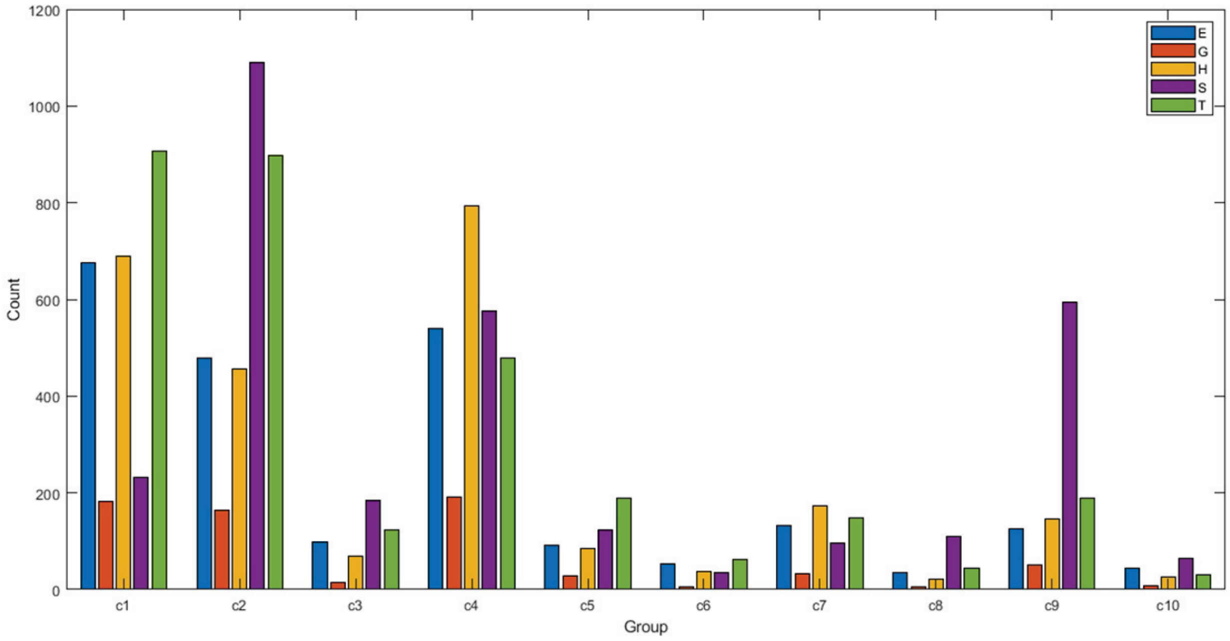

***Fig. 8**: A comparison of the grouping results for the real-time FM radio signal with reference to the mean of the correlation coefficient.*

features of the audio signals are extracted, namely MFCC, GTCC, entropy, and pitch. The feature correlations are then compared in each speech group to determine the threshold values of the correlation coefficients used for signal testing. The results confirm that the SFC algorithm concept can optimize threshold values and accurately provide sound signal validation while following the specified reference value.

Another interesting observation is that the distribution corresponds to the mean and standard deviation to determine the threshold value in the development step and is key to determining the proper threshold value in a real-time application to suit the nature and tone of the language used. Tests under actual case conditions yield satisfactory results since the SFC platform can operate quickly and responds to uninterrupted continuous operation. It should be noted that only five speech types are used as speech recognition clusters in real-time FM radio monitoring, thereby affecting the number of signals detected. Limitations include the broadcasting content, which may consist of long paragraphs (rather than specific words) and background voices or music, making classification more difficult. Even so, the results demonstrate the advantages and disadvantages of each feature implemented in an individual speech sound group. Furthermore, a specific feature coordination pattern would further improve system performance and precision.

REFERENCES

[1] F. Alías, J. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, May 2016, Art. no. 143.

[2] J.-H. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, no. 5, pp. 690–706, May 2011.

[3] L. Li, D. Wang, C. Zhang, and T. F. Zheng, "Improving short utterance speaker recognition by modeling speech unit classes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1129–1139, Jun. 2016.

[4] A. Jati and P. Georgiou, "Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1577–1589, Oct. 2019.

[5] L. Moro-Velaquez, E. Hernandez-Garcia, J. A. Gomez-Garcia, J. I. Godino-Llorente, and N. Dehak, "Analysis of the Effects of Supraglottal Tract Surgical Procedures in Automatic Speaker Recognition Performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 798–812, 2020.

[6] S. Ghaffarzadegan, H. Boril, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, Oct. 2016.

[7] J. Ming and D. Crookes, "Speech enhancement based on full-sentence correlation and clean speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 531–543, Mar. 2017.

[8] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, Dec. 2019.

[9] R. V. Sharan and T. J. Moir, "An overview of applications and advancements in automatic sound recognition," *Neurocomputing*, vol. 200, pp. 22–34, Aug. 2016.

[10] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[11] A. Pramanik and R. Raha, "Automatic speech recognition using correlation analysis," in *2012 World Congress on Information and Communication Technologies*, 2012, pp. 670–674.

[12] M. E. Rahaman, S. M. S. Alam, H. S. Mondal, A. S. Muntaseer, R. Mandal, and M. Raihan, "Performance analysis of isolated speech recognition technique using MFCC and cross-correlation," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019.

[13] R. Gupte, S. Hawa, and R. Sonkusare, "Speech Recognition Using Cross Correlation and Feature Analysis Using Mel-Frequency Cepstral Coefficients and Pitch," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020.

[14] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep learning based fusion approach for hate speech detection," *IEEE Access*, vol. 8, pp. 128 923–128 929, 2020.

[15] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204 951–204 962, 2020.

[16] O. Oriola and E. Kotze, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21 496–21 509, 2020.

[17] T. Juhana and S. Girianto, "An SDR-based multistation FM broadcasting monitoring system," in *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)*, 2017.

[18] M. A. B. Sahbudin, C. Chaouch, M. Scarpa, and S. Serrano, "IoT based song recognition for FM radio station broadcasting," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 2019.

[19] Y. Q. Lu, Q. N. Lu, D. Z. Chen, J. J. Yang, L. Zhang, and M. Huang, "FM broadcasting monitoring method based on time series analysis," in *Proceedings of the XXXIIIrd URSI General Assembly and Scientific Symposium (URSI GASS 2020)*, 2020.

[20] S. Potisuk, M. Harper, and J. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method,"

IEEE Transactions on Speech and Audio Processing, vol. 7, no. 1, pp. 95–102, 1999.

[21] N. Satravaha, P. Klinkhachorn, and N. Lass, "Tone classification of syllable-segmented Thai speech based on multilayer perceptron," in *Proceedings of the 35th Southeastern Symposium on System Theory*, 2003, pp. 392–396.

[22] S. Potisuk, "A system for extracting F0 contours of lexical tones using adaptive IIR notch filter with harmonic suppression," in *2016 International Conference on Asian Language Processing (IALP)*, 2016, pp. 116–119.

[23] S. Haykin, *Communication Systems*, 3rd ed. New York, USA: John Wiley & Sons, 1994.

[24] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.

[25] J.-M. Liu *et al.*, "Cough signal recognition with Gammatone Cepstral Coefficients," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 160–164.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[27] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006.

[28] N.-V. Vu, J. Whittington, H. Ye, and J. Devlin, "Implementation of the MFCC front-end for low-cost speech recognition systems," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2010, pp. 2334–2337.

[29] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[30] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Transactions on Signal Processing*, vol. 48, no. 6, pp. 1687–1694, Jun. 2000.

[31] L. Hui, B.-Q. Dai, and L. Wei, "A pitch detection algorithm based on AMDF and ACF," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 2006.

[32] 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*, 2009, pp. 33–36.

[33] S. Gonzalez and M. Brookes, "A Pitch Estimation Filter robust to high levels of noise (PEFAC)," in *2011 19th European Signal Processing Conference*, 2011, pp. 451–455.

[34] A. Ludloff, *Praxiswissen Radar und Radarsignalverarbeitung*, 3rd ed. Wiesbaden, Germany: Vieweg, 2002.

Narathep Phruksahiran received his Dipl.-Ing. degree in information engineering from University of the German Federal Armed Forces, Munich, Germany, in 2003 and Dr.-Ing. degree in electrical and information engineering from Chemnitz University of Technology, Chemnitz, Germany, in 2013. He is currently an Associate Professor in the Department of Electrical Engineering, Chulachomklao Royal Military Academy, Nakhon-Nayok, Thailand. His research interests are in cognitive radio, spectrum sensing, microwave radar remote sensing, electromagnetic scattering and radio wave propagation.