Received: April 14, 2022. Revised: July 21, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Mel-weighted Spectrogram Feature Extraction for Improved Speaker Recognition System

Yenni Astuti[1]* **Risanuri Hidayat[1]** **Agus Bejo[1]**

[1]*Electrical Engineering and Information Technology Department, Gadjah Mada University, Indonesia*
* Corresponding author's Email: yenni.stta@mail.ugm.ac.id

Abstract: Speaker recognition system is intended to recognize a person's identity. This task can be done by knowing the feature of the contained voice signals. The feature can be extracted using a feature extraction technique. One of the most popular feature extraction techniques is Mel-Frequency Cepstral Coefficient (MFCC). Until this time, the MFCC is still a challenging technique to be developed. In this work, a proposed technique is developed based on the MFCC. The original MFCC is modified in the process of filter bank and by adding spectrogram. The target of this work is to obtain a better recognition accuracy than the original MFCC. This target can be done by applying a technique called Mel-weighted spectrogram. The output of the proposed technique are spectrogram images which contain the feature of the voices. The spectrogram result is then classified using dissimilarity space based on Euclidean distance to identify the person's identity. For the dataset, this work uses 315 recorded voice signals, consisting of 3 speakers, each pronouncing five words repeatedly 21 times on seven different days. The performance of this system is evaluated by comparing the percentage of accuracy among the proposed technique, the original MFCC, and two other MFCC-based techniques. In this work, the proposed technique is better than the three other techniques with an accuracy up to 88.57%. From these results, the Mel-weighted spectrogram can be considered as a recommendation for obtaining a higher recognition rate in speaker recognition system.

Keywords: MFCC, Spectrogram, Mel-scale, Speaker recognition.

1. Introduction

A speaker recognition system is a part of speech processing field as well as speech recognition and language identification [1]. This system aims to recognize the identity of the speakers. From the type of the output, there are two categories of speaker recognition. The first type is known as speaker identification. The output of this system is the identity of the speaker. It can be name, age, or other aspects that are already stored in the database. The second one is known as speaker verification. The output of this system is a "Yes/No" form that verifies whether the speaker's is exactly the identity they claimed. Based on the spoken text, there are text-dependent and text-independent modes [2]. In the text-dependent mode, the spoken phrases are pre-determined. While in the text-independent mode, the spoken phrases can be anything. This work tries to build a speaker recognition system in speaker identity category with text-dependent mode.

Generally, recognition in machine learning can be made through two stages, known as the feature extraction process and the classification process. In the feature extraction process, the features contained in the input signals are extracted and stored in the system's database. A feature extraction process is the main phase of the speech recognition system [3]. This process gives a significant effect on system performance [4]. Meanwhile, in the classification process, the vectors output of the tested signal is computed to obtain a similarity value with the vectors stored in the database. This work focuses the discussion and experiment on the feature extraction technique as the process can effectively improve the recognition accuracy.

There are traditional feature extraction techniques that can be used for speaker recognition,

for example, Linear Prediction Cepstral (LPC) [5], Mel-Frequency Cepstral Coefficient (MFCC) [6], and Wavelet Transform [7]. In [5], three feature extraction technique is used, i.e., MFCC, LPC and ZCR. The combination of MFCC, LPC, and ZCR obtain an accuracy up to 92.8%. This feature extraction technique is, then, processed with ANN Artificial Neural Network) for the decision process. In [6], the study substituted the triangular filter bank of the MFCC with the Gaussian filter. In [7], the author used DWT (Discrete Wavelet Transform) combined with MFCC to extract the essential features of the voice samples. All these works show that the MFCC is one of the most used popular techniques and still have an open possibility for development. Considering its possibility, this work attempts to develop the MFCC to obtain a higher accuracy rate.

This paper proposes a new algorithm of feature extraction technique for speaker recognition system. The proposed algorithm of the speaker recognition is built based on the MFCC feature extraction technique. Our proposed work intends to improve the recognition accuracy of the MFCC. Because it is built from the MFCC concept, we analysed its effectiveness by comparing the proposed algorithm with the original MFCC, and two other MFCC-based techniques. The current MFCC is best applied in a noiseless environment, thus, this work evaluates the performance of the proposed algorithm in a noiseless environment. This algorithm was evaluated on a pre-built dataset corpus. The use of modified filter banks and spectrogram into the MFCC outperform the recognition accuracy.

Principally, the MFCC work similarly to how the ear perceive a sound [8, 9]. Inside the MFCC, i.e., in the Mel spectrum, there are a triangular filter bank and Mel-scale. This combination is the most important process in the MFCC. In this work, the triangular filter banks are substituted with Gaussian filter bank. The reason for its substitution is because Gaussian filter bank gives a smooth transition from one sub-band to the nearest sub-band while keeping its correlation [10]. After being processed in the filter bank and weighting by Mel-scale, in this work, spectrogram is applied to obtain the image of the voice power level scale. A spectrogram is a two-dimensional graph representing the time domain and frequency domain, with one more dimension, i.e., the pixel intensity that represents the signal amplitude. The spectrogram is a popular technique to recognize a person's emotion from his voice [11]. The combination of MFCC, Gaussian filter bank, and spectrogram is attempted to be applied in speaker recognition to improve the system's accuracy. Furthermore, the built system is evaluated using percentage of accuracy between the proposed algorithm, the original MFCC, and two other MFCC-based feature extraction technique.

For comprehensive description, this work is divided into several sections. Section 1 explains the general information of the speaker recognition system. Section 2 gives a brief description of the proposed algorithm, including the Gaussian filter bank and spectrogram explanation, along with the dataset used. Section 3 presents the experimental result and discussion. Finally, section 4 provides the conclusion of the experiment.

2. Methodology

The proposed algorithm in this work is built from the original MFCC. In the original MFCC, the processes consist of pre-emphasis, framing, Hamming windowing, FFT, filter bank & Mel-scale, and DCT (Discrete Cosine Transform). In the proposed algorithm, the triangular filter bank, which was used in the original MFCC, is substituted by the Gaussian filter bank. Furthermore, the DCT process of the original MFCC is substituted by the spectrogram. The three processes, i.e., FFT, Gaussian filter bank & Mel-scale, and spectrogram are then called Mel-weighted spectrogram. The block diagram of the proposed algorithm is shown in Fig. 1, which consist of six processes, i.e., pre-emphasis, framing, Hamming windowing, FFT, Filter bank & Mel scale, and spectrogram. For the classification, the Euclidean dissimalirity space is used.

2.1 Pre-emphasis

The system's input is a recorded voice. The digitalized voice is processed in a filter called a pre-emphasis filter. This filter produces the n^{th} output sample by subtracting the n^{th} input sample from the $(n\text{-}1)^{th}$ input sample on a scale of a [12]. The symbol a represents the pre-emphasis factor, and its value depends on its use. This pre-emphasis filter has a high pass response. Therefore, this filter blocks the DC offset of the input signal, which can significantly reduce the performance of feature extraction and classification modules. This filter also aims to even out the signal spectrum. It makes the signal more resistant to the effects of post-process cutting and improves classification precision and performance by minimizing the distance between the tested signal and the reference template data.

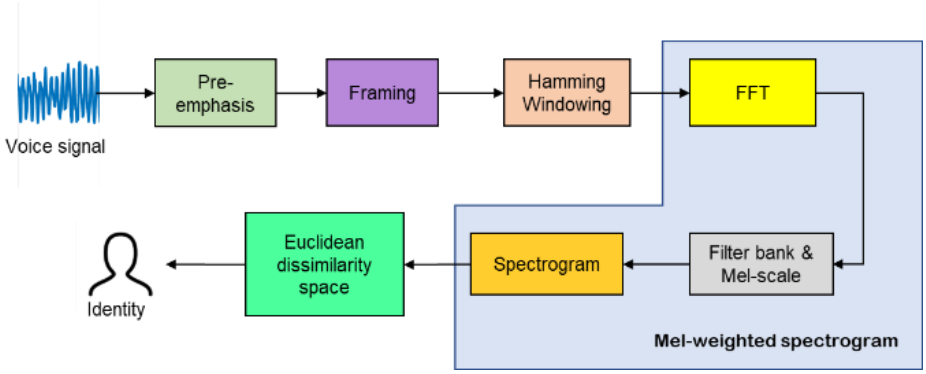

Figure. 1 Block diagram of Mel-weighted spectrogram system

In mathematical equation, the pre-emphasis process is shown in Eq. (1) [13].

$$y[n] = x[n] - ax[n-1] \quad (1)$$

where $y[n]$ is the voice signal after pre-emphasis, $x[n]$ is the recorded voice signal (in WAV format), a is a constant with a value between 0.9 and 1.0, (in this work set at 0.97), and n is the sampled signal with a value between zero and its frequency sampling, multiplied by signal time length) or written as $0 < n < (fs \times t)$.

After the pre-emphasis filter is completed, the voice signal is then normalized to obtain the same amplitude range for all the voice signals. In the normalization process, the amplitude of the n^{th} sampling voice signal is divided by the maximum absolute amplitude of the sampling signal. The mathematical formula for the normalization process is shown in Eq. (2) [6].

$$y_{norm}[n] = \frac{y[n]}{max|y(n)|} \quad (2)$$

with $y_{norm}[n]$ is a voice signal after normalized, and $y[n]$ is the output of the pre-emphasis process.

2.2 Framing

After the completion of the previous processing, i.e., pre-emphasis and normalization, the signal is then divided into several parts called frames. By nature, voice is a non-stationary signal. However, for a short duration, the voice signal can be assumed to be stationary. The voice parameters change approximately every 15 milliseconds [5]. The frame length commonly used in voice signal processing is 20 to 30 milliseconds [14]. Adjacent frames are set to overlap by 30% to 75% of the frame length. The frame size is a power of two to facilitate the use of FFT. Overlap between neighbourhood frames is applied to obtain a smooth transition and to capture the information that might be available in the frame boundary. In this work, the voice duration is one second, with a sampling frequency of 8,000 Hz. The length of each frame is set to 25 milliseconds with an overlap of 50% of each frame length or 12.5 milliseconds. Simply put, each frame contains 200 sampling signals, with an overlap of 100 sampling signals.

2.3 Hamming windowing

After several frames are built, each frame is subjected to a windowing process. This windowing process is carried out to maintain continuity in the border of the short-duration frames. This process must be applied when performing the Fourier transform using an FFT algorithm. One of the windowing techniques which is widely used and has good overall performance is Hanning window [15]. However, to obtain a smaller secondary lobe, the Hamming window is recommended [15]. Hence, in this paper, Hamming window is used.

Hamming window can be expressed mathematically as in Eq. (3) [16].

$$w[n] = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \quad (3)$$

with $w[n]$ is the Hamming windowing function and n is the sequence of the sampled signal whose value is between 0 and N–1. Variable N is the length of each frame. A voice signal after being multiplied by the window can be expressed as in Eq. (4) [16].

$$H[n] = y_{norm}[n] \times w[n] \quad (4)$$

with $H[n]$ is the resulting signal from the windowing, $y_{norm}[n]$ is the signals in each frame, $w[n]$ is the Hamming windowing function, as in Eq. (3), and n is the sampled signals whose value is between 0 and $(fs \times t)$.

2.4 Mel-weighted spectrogram

The Mel-weighted spectrogram consists of three sub-processes, i.e., Fast Fourier Transform (FFT), filter bank & Mel-scale, and spectrogram. The Fourier transform is applied to the windowing output signal using Discrete Fourier Transform (DFT). The DFT of the signal can be obtained using Eq. (5) [17].

$$X(k) = \sum_{n=0}^{N-1} x[n] exp\left(-j\frac{2\pi kn}{N}\right) \quad (5)$$

with k = 0, 1, 2, …, L–1, and L is the number of sampled signals, $X(k)$ is the complex value representing the magnitude and phase of the certain frequency component.

The Mel spectrum is computed by passing the Fourier transformed signal through a set of bandpass filters called the Mel-filter bank [18]. Mel is a unit of measure based on the frequency that is perceived by human ears. Mel spectrum is not as linear as the frequency spectrum of physical tones, as well as the human auditory system does not hear tones linearly. The Mel-scale is linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz [19]. Mel's approximation for the human ears' perceives frequency can be expressed as Eq. (6) [19].

$$f_{Mel} = 2595 \times \log_{10}(1 + \frac{f}{700}) \quad (6)$$

with f represents the frequency as measured in Hertz, and f_{Mel} denotes the frequency perceived by human ears represented by Mel-scale.

Filter banks can be implemented both in time domain and frequency domain. In MFCC computation, filter banks are generally implemented in the frequency domain. The center frequency of the filter is, normally, evenly spaced on the frequency axis. However, to simulate the perception of the human auditory system, a nonlinear function, as expressed in Eq. (6), is used. The number of filter banks used in the Mel spectrum is 20 – 40 filters [20]. Commonly, the triangular filter is used as filter banks in MFCC. This triangular shape filter is a rough estimation of the actual frequency response of the basilar membrane, which peaks at a certain frequency but also responds to the surrounding frequency band [21].

The Mel spectrum of the magnitude spectrum, $X(k)$, is processed by multiplying the magnitude spectrum of each triangular Mel weighting filter, as seen in Eq. (7) [18].

$$s(m) = \sum_{k=0}^{N-1}[|X(k)|^2 H_m(k)] \quad (7)$$

with m is a value between 0 and M–1, and M is the number of the triangular filter bank, $H_m(k)$ is the weight given to the k^{th} energy spectrum bin that affect the m^{th} output band as express as in Eq. (8) [18].

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (8)$$

With the value of n varies between 0 to M –1.

2.5 Gaussian filter

In this work, the triangular filters as the filter banks in the MFCC is substituted with the Gaussian filter banks. There are several reasons for substituting the triangular filter bank with a Gaussian filter bank [10]. The first reason, this filter produces smoother transitions from one sub-band to another by maintaining its correlation. The second reason, the means and variances of the Gaussian filter can be changed to control the overlap of the neighbouring sub-bands. The third, these filter design parameters can be calculated very easily from the midpoints and endpoints which lie at the base of the triangular filter on the original MFCC. The Gaussian filter can be applied using Eq. (9).

$$\psi_i{}^g = e^{-\frac{(k-k_{bi})^2}{2\sigma_i^2}} \quad (9)$$

with k_{bi} represents the point between i^{th} transfer limit, and σ_i is the i^{th} standard deviation. The standard deviation value can be calculated using Eq. (10).

$$\sigma_i = \frac{k_{bi+1}-k_{bi}}{\alpha} \quad (10)$$

with α represents variance control parameter. In this work, α is set to two since this value is recommended in [10].

2.6 Spectrogram

The spectrogram is a graph in two dimensions, with one axis representing frequency, and the other axis representing time, plus one dimension representing pixel intensity [22]. The color of each point on the graph represents the signal's amplitude at a certain frequency and time. The spectrogram is

considered a very accurate representation of audio information [23]. Before the spectrogram process is performed, Mel spectra are represented on a log scale. This technique produces a signal in the cepstral domain with a frequency peak related to the tone signal and a formant number representing the low-frequency peak.

2.7 Classification

For the decision process, classification is applied to obtain the speaker's identity whose voice is tested. It is performed by comparing the extracted features of the tested voice with all stored voice models from the training stage.

Various classification techniques can be used, such as Hidden Markov Model (HMM), Support Vector Machine (SVM), and dissimilarity space. In this work, a dissimilarity space classification technique based on Euclidean distance is selected. Euclidean distance is the simplest and the most effective method for comparing two vectors. The Euclidean distance is mathematically shown in Eq. (11)[24].

$$d(a,b) = \sqrt{\sum_{k=1}^{n}[a_k - b_k]^2} \qquad (11)$$

With a is a set of points in vector a which consists of $[a_1, a_2, \ldots, a_n]$ and b is a set of points in vector b which consists of $[b_1, b_2, \ldots, b_n]$. The size of the two vectors must be the same. Euclidean distance is very effective for small amounts of data [25].

2.8 Dataset

The main purpose of a speaker recognition system is to recognize a person's identity by their voice regardless its language and words. The proposed system can be applied for any speaker's languages depend on the dataset availability. In this work, the pre-built dataset utilizes Indonesian speaker's voices. For additional information, the words that were spoken by each participant are "Aku, "Saya", "Dan", "Tidak", and "Nggak". These five words were chosen because it is the most frequently spoken words in the Indonesian language. These words are repeated 21 times in seven days. The difference days are used to facilitate the possibility of voice variation from the speakers. The total data used is 315 voice signals from three speakers. The speakers consist of a man and two women of the same age. From the total data, the 215 voice signals are used for training, and the other 105 voice signals are used for testing. The 315 voice signals are saved in mono WAV form, with a sampling frequency of 8,000 Hz. The duration of each voice signals is set to one second long.

3. Results and discussion

In the training stage, the 210 training voice signals are processed in the pre-emphasis filter. The results are then divided into several frames. As the duration of each signal is one second with a frequency sampling 8,000 Hz, each frame then contains 200 sampled signals. Each frame is overlapped 50% or 100 sampled signals. The number of frames which is generated from this process is 79 frames. These frames are then processed by Hamming windowing. The windowing result is then transformed to the frequency domain using the FFT algorithm. The result of this process is a single-side frequency spectrum for each frame voice signals. The results of the FFT are then processed in the filter bank, which uses a Gaussian filter bank. After being processed through the filter bank, the FFT spectrum is subjected to a Mel scale to simulate the human auditory perception. Fig. 2 shows the Gaussian filter bank after Mel scale. The axis of the filter shape is frequency (Hz), and the ordinate is the amplitude.

The results of the Mel scale are Mel coefficient, further converted into a spectrogram image. Fig. 3 to Fig. 5, respectively, show the example of spectrograms of the first participant (P1), the second participant (P2), and the third participant (P3) when pronouncing "*aku*". The axis of each spectrogram is time (in second), and the ordinate is the frequency (in Hertz). After the spectrogram of all the training data is stored in the database, the next process is testing.

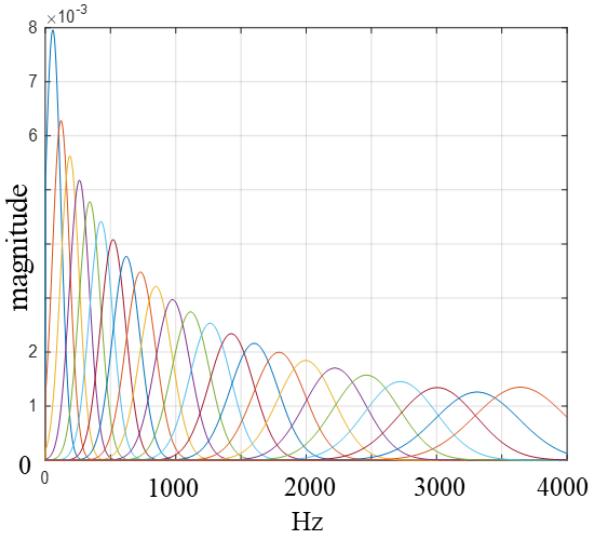

Figure. 2 Gaussian filter banks and Mel-scale

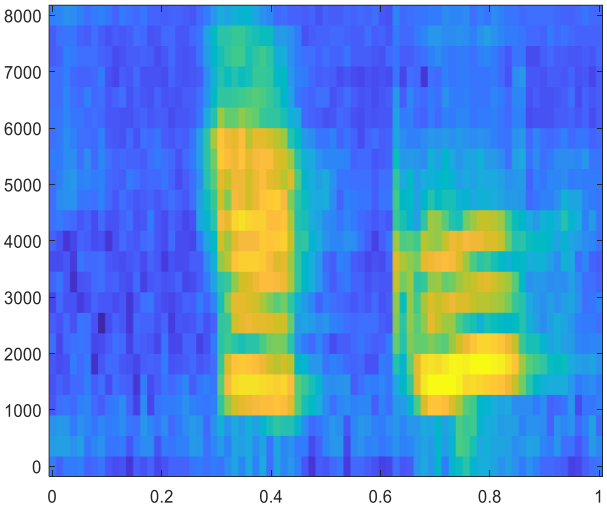

Figure. 3 Spectrogram of the participant-1 (P1) pronounces a word "Aku"

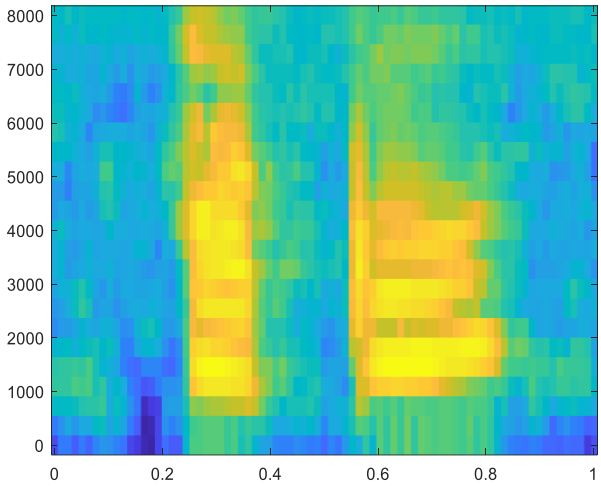

Figure. 4 Spectrogram of the participant-2 (P2) pronounces a word "Aku"

In the testing process, 105 voice signals are tested one by one to obtain the recognition results, for example, the voice signal namely U1 is input to the system. In the system, this file was subjected to pre-emphasis filter, framing, Hamming windowing, FFT, filter bank and Mel scale, and spectrogram, like the training process. The difference is the computation of the input toward the database. The spectrogram of signal U1 is then compared to all database's spectrogram to obtain the smallest dissimilarity space. The training signal with the smallest dissimilarity value compared to the tested signal becomes the recognition result. Table 1 and Table 2, respectively, show the recognition result for the original MFCC and the Mel-weighted spectrogram, using 30 filter banks (NF=30) for 20 tested data.

In Table 1, for the 20 tested signal, data U3, U6, U16, U19, and U20 result in wrong recognition. The

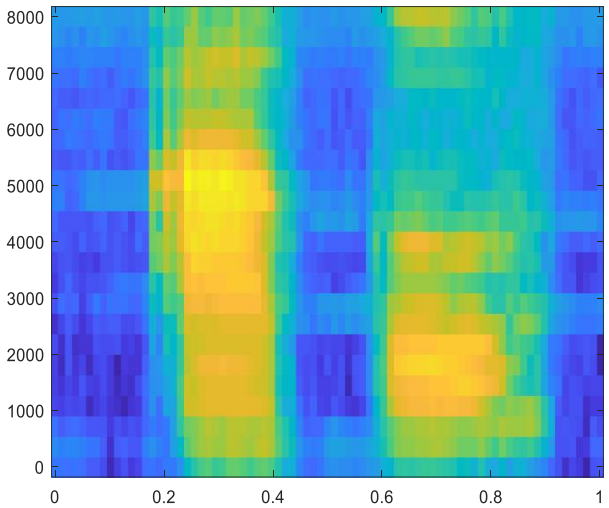

Figure. 5 Spectrogram of the participant-3 (P3) pronounces a word "Aku"

data are expected to be belonged to Participant 1 (P1), but the system recognize them as Participant 2 (P2). However, in Table 2, where the Mel weighted spectrogram is used, data U3, U6, U16, U19, and U20 are recognized correctly, i.e., as P1. The other 85 tested data are not shown in Table 2 and Table 3 but summarized as the percentage accuracy graph in the Fig. 6.

Table 1. Recognition result of original MFCC for 20 tested data

Filename	Participant ID	Recognized as
"U1"	P1	P1
"U2"	P1	P1
"U3"	P1	P2
"U4"	P1	P1
"U5"	P1	P1
"U6"	P1	P2
"U7"	P1	P1
"U8"	P1	P1
"U9"	P1	P1
"U10"	P1	P1
"U11"	P1	P1
"U12"	P1	P1
"U13"	P1	P1
"U14"	P1	P1
"U15"	P1	P1
"U16"	P1	P2
"U17"	P1	P1
"U18"	P1	P1
"U19"	P1	P2
"U20"	P1	P2

Table 2. Recognition result of the Mel-weighted spectrogram for 20 tested data

Filename	Participant ID	Recognized as
"U1"	P1	P1
"U2"	P1	P1
"U3"	P1	P1
"U4"	P1	P1
"U5"	P1	P1
"U6"	P1	P1
"U7"	P1	P1
"U8"	P1	P1
"U9"	P1	P1
"U10"	P1	P1
"U11"	P1	P1
"U12"	P1	P1
"U13"	P1	P1
"U14"	P1	P1
"U15"	P1	P1
"U16"	P1	P1
"U17"	P1	P1
"U18"	P1	P1
"U19"	P1	P1
"U20"	P1	P1

The evaluation of the system is performed by comparing the recognition accuracy of the proposed algorithm with the other feature extraction techniques in various filter bank numbers. Because built from the MFCC concept, the proposed system is evaluated by comparing it with the original MFCC, and two others MFCC-based, i.e., Gaussian MFCC [6], and DWT-MFCC [7]. Inside the original MFCC, the system's process consists of pre-emphasis, framing, windowing, FFT, filter bank & Mel-scale, and DCT. In the Gaussian MFCC, the filter bank block of the original MFCC is modified using Gaussian filter bank. In the DWT-MFCC, the *Discrete Wavelet Transform* (DWT) is applied before the MFCC process. The recommendation wavelet used for the system was Biorthogonal 2.2 level 2. Furthermore, in the proposed system, the filter bank and the DCT of the original MFCC, respectively, is substituted with the Gaussian filter bank and Spectrogram.

Fig. 6 shows the percentage bar chart of the Mel-weighted spectrogram compared to the original MFCC, Gaussian MFCC, and DWT-MFCC in various filter banks number (*NF* = 20, 25, 30, 35, 40). From the bar chart in the Fig. 6, for *NF* = 20, the original MFCC result recognition accuracy 74.29%. This rate is lower than the Gaussian MFCC with accuracy 78.10%, and the DWT-MFCC with accuracy 78.10%. The highest accuracy for this scenario (*NF* = 20) is the Mel-weighted spectrogram with accuracy 88.57%. For filter numbers equal to 25 (*NF* = 25), the DWT-MFCC results the lowest rate with accuracy 76.19% compared to the Gaussian MFCC (accuracy 79.05%), and original MFCC (accuracy 80%). The highest accuracy is obtained from the Mel-weighted spectrogram with the accuracy 82.86%. For *NF* = 30, the proposed system results a highest accuracy with recognition 80% compared to the original MFCC (77.14%), the DWT-MFCC (73.33%), and the Gaussian MFCC (79.05%).

For *NF* = 35, the Mel-weighted spectrogram shows the highest accuracy with the recognition rate 84.76% compared to the original MFCC and DWT-MFCC both with the accuracy 76.19%, and the Gaussian MFCC with the accuracy 77.14%. The last filter numbers, i.e., *NF* = 40, the proposed technique result the highest accuracy with accuracy 88.57% compared to the original MFCC and Gaussian MFCC, both, with accuracy 76.19%, and the DWT-MFCC with accuracy 78.10%. In this experiment, it is shown that the Mel-weighted spectrogram offers a better recognition accuracy with percentage up to 88.57% that occurs in filter numbers equal to 20 and 40.

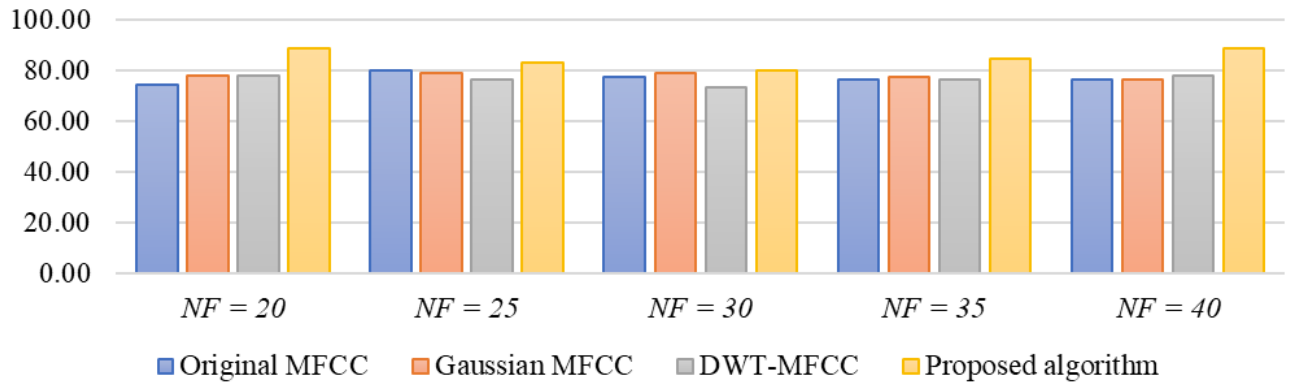

Figure. 6 Percentage accuracy of the proposed method, original MFCC, Gaussian MFCC, and DWT-MFCC

Gaussian	P1	P2	P3
P1	94%	3%	3%
P2	6%	91%	3%
P3	11%	9%	80%

Figure. 7 Confusion matrix of Mel-weighted spectrogram recognition

Additionally, because the proposed system has the highest result when the *NF* = 40, the confusion matrix from the speaker recognition system is presented with the *NF* = 40 as in Fig. 7. In Fig. 7, the highest correct recognition is obtained from Participant 1 with the value of 94%, and the lowest correctly recognition is obtained from Participant 3 with the value of 80%.

4. Conclusion

Ideally, a speaker recognition system can recognize a person's identity by their voice. To build the system, feature extraction is needed to obtain some special features contained in the voice signals. There are many features extraction techniques that can be considered in the speaker recognition system. One of the popular techniques is Mel Frequency Cepstral Coefficient (MFCC). This technique applies Mel-scale in the spectrum to imitate human auditory system.

In this work some of the MFCC processes are modified. The first modification is the implementation of the Gaussian filter banks in the original MFCC instead of the triangular filter banks. The second modification is the use of spectrogram after the filter bank & Mel-scale process. The MFCC process from pre-emphasis until filter banks & Mel-scale is applied to the system then followed by a spectrogram process. In other words, the feature extraction output is no longer a Mel coefficient, but images that contains its histogram. This histogram is then stored as a database to be used for classification or decision process. Both modifications are intended to obtain a higher recognition rate.

From the experiment results, the proposed algorithm (Mel-weighted spectrogram) performs a higher accuracy compared to the original MFCC, Gaussian MFCC, and DWT-MFCC with the accuracy up to 88.57%. Based on the confusion matrix, the proposed system obtains the highest recognition rate at 94% which results from recognizing Participant 1, and the lowest rate at 80%, which results from recognizing Participant 3. For further work, it is suggested to add more speaker participants and speakers' language in the system's dataset, also to add some noise to deal with a real environment.

Conflicts of Interest

The authors declare that we do not have any circumstances or interest that may affect the results discussed in this manuscript.

Author Contributions

Conceptualization, Yenni Astuti and Risanuri Hidayat; methodology, Yenni Astuti and Risanuri Hidayat; software, Yenni Astuti and Agus Bejo; validation, Risanuri Hidayat and Agus Bejo; formal analysis, Yenni Astuti, Risanuri Hidayat and Agus Bejo; investigation, Yenni Astuti; resources, Yenni Astuti; data curation, Yenni Astutio; writing—original draft preparation, Yenni Astuti; writing—review and editing, Risanuri Hidayat, Agus Bejo, and Yenni Astuti; visualization, Yenni Astuti; supervision, Risanuri Hidayat and Agus Bejo; project administration, Yenni Astuti; funding acquisition, Yenni Astuti.

References

[1] J. P. Campbell, "Speaker Recognition: A Tutorial", In: *Proc. of the IEEE*, Vol. 85, No.9, pp. 1437-1462, 1997.

[2] A. Irum and A. Salman, "Speaker verification using deep neural networks: A review", *International Journal of Machine Learning and Computing*, Vol. 9, No. 1, pp. 20-25, 2019.

[3] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition", In: *Proc. of Speech and Language Processing for Human-Machine Communicatio*n, Vol. 664, pp. 195-207, 2018.

[4] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network", *IEEE Access*, Vol. 7, pp. 125868-125881, 2019.

[5] N. Chauhan, T. Isshiki, and D. Li, "Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database", In: *Proc. of IEEE 4th International*

Conf. on Computer and Communication Systems, pp. 130-133, 2019.

[6] Y. Astuti, R. Hidayat, and A. Bejo, "Feature Extraction using Gaussian-MFCC for Speaker Recognition System", In: *Proc. of 5th International Conf. Information Technology, Information System and Electrical Engineering*, pp. 186-190, 2021.

[7] F. Amelia and D. Gunawan, "DWT-MFCC Method for Speaker Recognition System with Noise", In: *Proc. of 7th International Conf. on Smart Computing and Communications*, pp. 1-5, 2019.

[8] 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 Systems*, Vol. 13, No. 2, pp. 74-82, 2020, doi: 10.22266/ijies2020.0430.08.

[9] K. P. Bharath and M. R. Kumar, "ELM Speaker Identification for Limited Dataset Using Multitaper based MFCC and PNCC Features with Fusion Score", *Multimedia Tools and Applications*, Vol. 79, No. 39, pp. 28859-28883, 2020.

[10] S. Singh and E. G. Rajan, "Application of Different Filters in Mel Frequency Cepstral Coefficients Feature Extraction and Fuzzy Vector Quantization Approach in Speaker Recognition", *International Journal of Engineering Research & Technology*, Vol. 2, No. 6, pp. 3171-3182, 2013.

[11] H. Kaya, A. A. Salah, A. Karpov, O. Frolova, A. Grigorev, and E. Lyakso, "Emotion, Age, and Gender Classification in Children's Speech by Humans and Machines", *Computer Speech & Language*, Vol. 46, pp. 268-283, 2017.

[12] S. M. Qaisar, "Isolated Speech Recognition and Its Transformation in Visual Signs", *Journal of Electrical Engineering & Technology*, Vol. 14, No. 2, pp. 955-964, 2019.

[13] R. Hidayat and A. Winursito, "A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition", *International Journal of Intelligent Engineering Systems*, Vol. 14, No. 1, pp. 12-21, 2021, doi: 10.22266/ijies2021.0228.02.

[14] S. B. Dhonde and S. M. Jagade, "Mel-Frequency Cepstral Coefficients for Speaker Recognition: A Review", *International Journal of Advance Engineering and Research Development*, Vol. 2, No. 5, pp. 1115-1119, 2015.

[15] S. D. Fassois, "Identification, Model-Based Methods", *In Encyclopedia of Vibration*, Vol. 2, S. Braun, UK: Elsevier, pp. 673-685, 2002.

[16] R. Hidayat and A. Winursito, "Improving Accuracy of Isolated Word Recognition System by Using Syllable Number Characteristics", *International Journal of Technology*, Vol. 11, No. 2, pp. 411-421, 2020.

[17] K. J. Devi, N. H. Singh, and K. Thongam, "Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network", *Microprocessors and Microsystems*, Vol. 79, p. 103264, 2020.

[18] K. S. Rao and K. E. Manjunath, "Appendix A: MFCC Features", *In Speech Recognition Using Articulatory and Excitation Source Features*, pp. 85-88, 2017.

[19] Y. Astuti, R. Hidayat, and A. Bejo, "Comparison of Feature Extraction for Speaker Identification System", In: *Proc. of 3rd International Seminar on Research of Information Technology and Intelligent Systems*, pp. 642-645, 2020.

[20] A. Revathi, C. Ravichandran, P. Saisiddarth, and G. S. R. Prasad, "Isolated Command Recognition Using MFCC and Clustering Algorithm", *SN Computer Science*, Vol. 1, No. 2, pp. 1-7, 2020.

[21] R. Singh, *Profiling Humans from their Voice. Springer*, p. 180, 2019.

[22] L. Nanni, A. Rigo, A. Lumini, and S. Brahnam, "Spectrogram Classification Using Dissimilarity Space", *Applied Science*, Vol. 10, No. 12, pp. 1-17, 2020.

[23] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram Based Multi-Task Audio Classification", *Multimedia Tools and Applications*, Vol. 78, No. 3, pp. 3705-3722, 2019.

[24] R. Bu, "Spectrogram, Embedded System and Speech Recognition", In: *XIII International PhD Workshop (OWD)*, pp. 277-280, 2011.

[25] M. Mohibullah, M. Z. Hossain, and M. Hasan, "Comparison of Euclidean Distance Function and Manhattan Distance Function Using K-Mediods", *International Journal of Computer Science and Information Security*, Vol. 13, No. 10, pp. 61-71, 2015.