https://doi.org/10.31436/iiumej.v23i1.1760

WAVELET DETAIL COEFFICIENT AS A NOVEL WAVELET-MFCC FEATURES IN TEXT-DEPENDENT SPEAKER RECOGNITION SYSTEM

SYAHRONI HIDAYAT[1,2*], MUHAMMAD TAJUDDIN[3], SITI AGRIPPINA ALODIA YUSUF[2], JIHADIL QUDSI[4], NENET NATASUDIAN JAYA[5]

[1]*Dept. of Agricultural Engineering, University of Mataram, Mataram City, Indonesia*
[2]*Research and Development, Sekawan Institute, Mataram City, Indonesia*
[3]*Dept. of Computer Science, Universitas Bumigora, Mataram City, Indonesia*
[4]*Dept. of Medical Record, Politeknik Medica Farma Husada, Mataram City, Indonesia*
[5]*Dept. of Management, Universitas Mahasaraswati Mataram, Mataram City, Indonesia*

**Corresponding author: syahronihidayat@unram.ac.id*

(Received: 23rd December 2020; Accepted: 12th August 2021; Published on-line: 4th January 2022)

ABSTRACT: Speaker recognition is the process of recognizing a speaker from his speech. This can be used in many aspects of life, such as taking access remotely to a personal device, securing access to voice control, and doing a forensic investigation. In speaker recognition, extracting features from the speech is the most critical process. The features are used to represent the speech as unique features to distinguish speech samples from one another. In this research, we proposed the use of a combination of Wavelet and Mel Frequency Cepstral Coefficient (MFCC), Wavelet-MFCC, as feature extraction methods, and Hidden Markov Model (HMM) as classification. The speech signal is first extracted using Wavelet into one level of decomposition, then only the sub-band detail coefficient is used as the feature for further extraction using MFCC. The modeled system was applied in 300 speech datasets of 30 speakers uttering "HADIR" in the Indonesian language. K-fold cross-validation is implemented with five folds. As much as 80% of the data were trained for each fold, while the rest was used as testing data. Based on the testing, the system's accuracy using the combination of Wavelet-MFCC obtained is 96.67%.

ABSTRAK: Pengecaman penutur adalah proses mengenali penutur dari ucapannya yang dapat digunakan dalam banyak aspek kehidupan, seperti mengambil akses dari jauh ke peranti peribadi, mendapat kawalan ke atas akses suara, dan melakukan penyelidikan forensik. Ciri-ciri khas dari ucapan merupakan proses paling kritikal dalam pengecaman penutur. Ciri-ciri ini digunakan bagi mengenali ciri unik yang terdapat pada sesebuah ucapan dalam membezakan satu sama lain. Penyelidikan ini mencadangkan penggunaan kombinasi Wavelet dan Mel Frekuensi Pekali Cepstral (MFCC), Wavelet-MFCC, sebagai kaedah ekstrak ciri-ciri penutur, dan Model Markov Tersembunyi (HMM) sebagai pengelasan. Isyarat penuturan pada awalnya diekstrak menggunakan Wavelet menjadi satu tahap penguraian, kemudian hanya pekali perincian sub-jalur digunakan bagi pengekstrakan ciri-ciri berikutnya menggunakan MFCC. Model ini diterapkan kepada 300 kumpulan data ucapan daripada 30 penutur yang mengucapkan kata "HADIR" dalam bahasa Indonesia. Pengesahan silang K-lipat dilaksanakan dengan 5 lipatan. Sebanyak 80% data telah dilatih bagi setiap lipatan, sementara selebihnya digunakan sebagai data ujian. Berdasarkan ujian ini, ketepatan sistem yang menggunakan kombinasi Wavelet-MFCC memperolehi 96.67%.

KEY WORDS: *Discrete wavelet transforms, Feature extraction, Hidden Markov models, Speaker recognition, Wavelet coefficients.*

https://doi.org/10.31436/iiumej.v23i1.1760

1. INTRODUCTION

Speaker recognition is the process of recognizing the speaker from his speech by comparing the sound biometrics of the words he has spoken with his pronunciation model, which has previously been used as a reference and stored in a database [1]. Speaker recognition can be used in many aspects; this technology is expected to make daily life easier by accessing personal devices remotely. As a biometric tool, speaker recognition also can be applied in secure access voice control, information structuring, customizing services, and forensic investigation. As a non-invasive biometric, speech can be collected without the knowledge of the speaker [2]. The speaker recognition system first appeared in 1962, in an article published by Nature entitled Voiceprint Identification. The research claimed had developed a method for identifying an individual with high success rates. The research developed a visual representation of speech called a spectrogram. However, the first successful implementation in speaker recognition was a text-dependent system developed in 1977, using Euclidean distance as a verification decision. In the last several years, many researchers have been interested in this field and made significant improvements to it [3], especially in feature extraction and classification.

In speaker recognition, feature extraction is a necessary process that can affect the performance of the system recognition. These features are used to represent and describe the signal. Each speaker has a unique vocal characteristic; this leads to two different features: morphological and behavioral features. Morphological features are determined by the length of the vocal tract and the size of vocal folds. At the same time, the behavioral features are determined by education, background, parental influence, personality type, place of birth, and language [1]. There are a few properties of good features, such as the ease of measurement and extraction, robustness, uniqueness, and independence between one feature and another.

Several feature extraction techniques are commonly used in speaker recognition, such as the Mel Frequency Cepstral Coefficient (MFCC), Gammatone Frequency Cepstral Coefficient (GFCC), Wavelet, and the combination of Wavelet-MFCC. Each of the methods has its advantages and disadvantages. Feature extraction using MFCC is found in [4], the research employed MFCC to extract the features. MFCC is an excellent features extractor; it can form features to mimic human hearing [5]. Although it was said that MFCC is the best feature extractor, especially in distant-talking speaker recognition [1], in the matter of robustness towards noise, GFCC is better [6]. Ayoub et al. in [7] employed GFCC as a features extractor in speaker identification system over VoIP network. The Discrete Wavelet Transformation (DWT) has also become popular and is employed in a variety of applications. The extension version of DWT, Wavelet Packet Transform (WPT), has superior presentation to DWT; it gives more flexibility in decomposition since WPT decomposes both details and approximations [8]. It is stated that the features extracted using a wavelet give equal accuracy to MFCC. Aside from that, many previous kinds of research claim that extracting features using Wavelet yielded better results compared to MFCC [9]. WPT was employed in [10] to extract features on TIMIT and TALUNG databases. However, it is shown that in the speaker recognition system, MFCC is commonly used as a feature extractor and can be combined with another new method such as DNN [1].

Several previous types of research have combined Wavelet and MFCC, forming MFCC based on Wavelet to overcome the noisy environment. Application of both, in combination, integrates the merits of both methods. MFCC can represent the speech spectrum in a compact form and is based on the model of human auditory perception, but the FFT process can cause the loss of information [11], [5]. Meanwhile, the Wavelet can map the signal into the time and

https://doi.org/10.31436/iiumej.v23i1.1760

frequency domain; thus, no information is lost in this process [9]. In [12], MFCC was combined with 3 level decomposition of DWT to extract signal uttering digits 0-9 in the Indian language. WPT-MFCC was combined in [13], employing all of the sub-band obtained from WPT to find the MFCC. While in [14], MFCC combined with Wavelet sub-band coefficient (SBC) in an isolated word for speaker recognition. The combination of Wavelet-MFCC yielded better results than conventional MFCC and GMM, as stated in [15].

The modeling and decision making in speaker recognition also vary, starting from the widely used HMM found in [4], [16], vector quantization (VQ) found in [9], SVM found in [14], the classical technique Gaussian Mixture Model (GMM) found in [10], [13], [12], and ANN found in [17] and [18]. However, from the decision-making model, HMM is more suitable for modeling speaker features, especially in text-dependent speaker recognition systems, where there are two types of HMM, Ergodic and left-right [19].

Much previous research that employed Wavelet-MFCC as feature extraction either used approximated sub-band coefficient as found in [9], or a combination of approximated and detailed sub-band of the sub-band, as found in [12], [17]. There are three aims in this research. First, we determine the best sub-band coefficient as an input for another feature extraction using MFCC. Second, we determine the best wavelet family in Wavelet-MFCC feature extraction method. We also try to find the influence of gender in the recognition system. Then, we compare the proposed method with other feature extractions such as conventional MFCC, MFCC + delta, and MFCC + delta-delta. As for the decision-making, we employed HMM.

2. RELATED WORKS

As the algorithm evolves for speaker recognition, one of the best algorithms is Mel Frequency Cepstral Coefficient (MFCC) [20]. Many previous researchers have used this algorithm to extract features from a sound signal. As in [21], MFCC was employed as a features extractor in the NTT database, and the Japanese Newspaper Article Sentences (JNAS) database and Vector Quantization (VQ) were used as a recognizer. The obtained result shows excellent accuracy, which is 98.75%.

It is said that GFCC works better than MFCC to extract features from noisy signals and shows an encouraging result. [7] studied and evaluated the performance of the method in text-independent over VoIP networks and obtained high identification rates. [10] applied WPT to extract features in two public databases, TALUNG and TIMIT. The research proposed a new method based on a multilayer neural network to reduce the dimension of wavelet packet features. The recognition rate obtained was 94% in the TALUNG database. WPT also applied in [11], the research studied speaker identification for security systems based on the energy of speaker utterance and achieved a 96.6% of recognition rate.

The hybrid features were based on Wavelet-MFCC proposed in [12]. The features were extracted in ten isolated Hindi digits using 3 level decomposition of Wavelet, then extracted using MFCC on every sub-band. The methods yielded 100% average performance. Adam et al. [22] proposed an improved feature extraction method called Wavelet Cepstral Coefficient (WCC) for isolated speech recognition in the English alphabet. The research replaced the DFT with DWT in order to obtain the merit of the wavelet transform. The coefficient from DWT is then used as a feature vector to calculate the log power spectrum and DCT.

https://doi.org/10.31436/iiumej.v23i1.1760

3. METHODOLOGY

In this research, generally, the proposed method consists of four steps: collecting database, preprocessing, feature extraction, and speech recognition. Fig. 1 shows the steps of the proposed method.

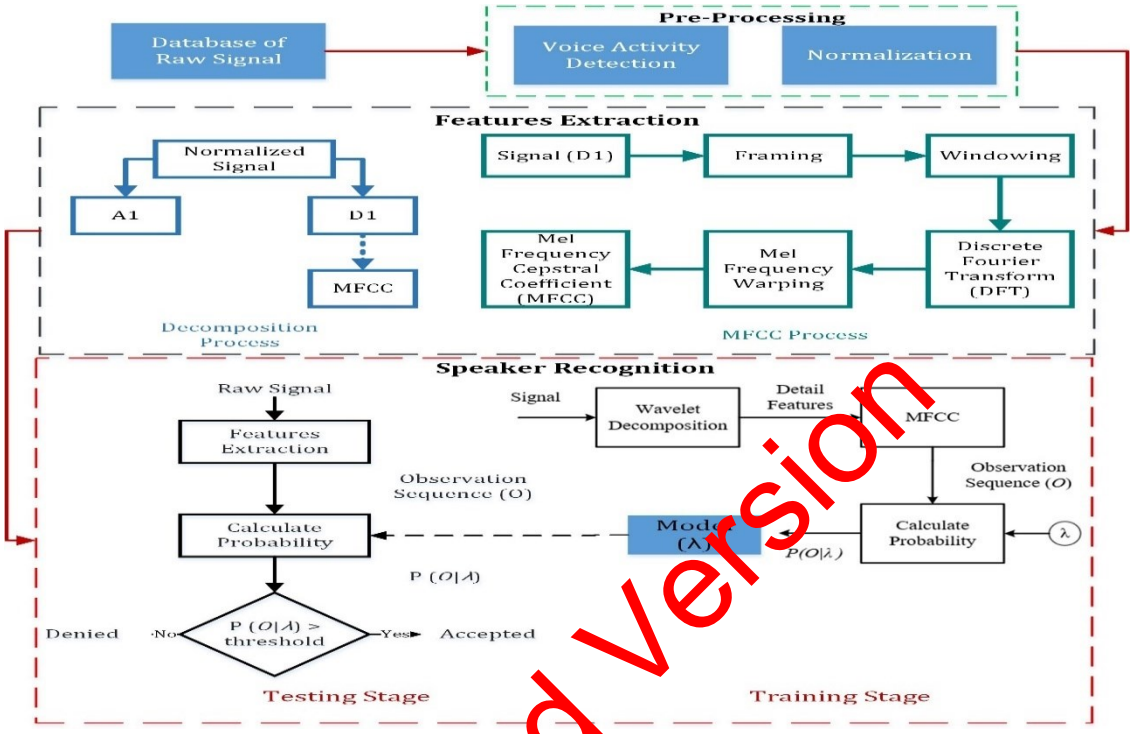

Fig. 1 Proposed Method

3.1. Database

The database used in this research was created with the help of 30 volunteers consisting of 20 male and ten female adults. All of the speakers were native Indonesian speakers and asked to utter the word "HADIR" in Indonesian, which in English means "PRESENT," and repeat the word ten times. The total dataset obtained was 300 datasets. In accordance with collecting the sound database, this research used a smartphone integrated with a microphone and used Audacity software. There were two settings done in recording; the first was an environmental setting, and the second was the recording properties setting in Audacity software. In the environmental setting, the recording was taken indoors to reduce the noise, and the recorder was placed 0.5 meters from the speaker; the Audacity setting was shown in the following table.

Table 1: Recording Properties

No	Variable	Value
1	Recording channel	Mono
2	Frequency Sampling	8096
3	Bits per Sample	16-bit PCM
4	Format	*.wav

3.2. Preprocessing

The aim of preprocessing is to enhance the quality of the signal. There were three steps in the preprocessing stage that consisted of noise reduction, voice activity detection, and signal

https://doi.org/10.31436/iiumej.v23i1.1760

normalization. In the first step, we applied a pre-emphasis filter to reduce the noise. Pre-emphasis is also able to emphasize the high frequency in the signal. Generally, the coefficient used in pre-emphasis oscillates between 0.9-0.97. In this research, we used $\alpha = 0.97$. Therefore, the following Eq. (1) is used to calculate pre-emphasis, where y[n] is the output, x[n] is the input signal n, and x[n-1] is the previous signal.

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

The next step was applying the voice activity detection algorithm. By applying this algorithm, the silent sound in the signal was removed [23]. This algorithm was developed by [23], and this algorithm uses the value of energy *E* as a threshold to distinguish the silent sound in the signal. The energy of the signal can be calculated using Eq. (2).

$$E = \frac{1}{N}\sum_{n=1}^{N}|x(n)|^2 \quad (2)$$

The last step of preprocessing was signal normalization, S_{norm}. Signal normalization aims to obtain the same magnitude value. By doing this, the maximum magnitude value is ±1. Eq. (3) was used to obtain the normalized signal. S_{norm} is the normalized signal, and $max|S|$ is the maximum value of the signal.

$$Snorm = \frac{s}{max\ |S|} \quad (3)$$

3.3. Feature Extraction

Feature extraction is a process to extract features from the signal. The extracted features must be unique and have high variability, among other features. In this step, two different methods were combined, wavelet and MFCC. The goal of this combination was to take advantage of both methods.

For wavelet feature extraction, this research applied Discrete Wavelet Transformation (DWT). Wavelet transformation is suitable for analyzing stationary and non-stationary signals since it can localize the signal in the time-frequency domain and has multi-resolution characteristics [12]. The signal then decomposed into one level decomposition. The decomposition itself was a convolution and decimation process of a signal by a factor of two [24], the result was two sub-bands for each level, sub-band A1, also known as approximation coefficient, obtained from low pass filter (LPF) g[n] and sub-band D, known as detail coefficient, obtained from high pass filter (HPF) h[n]. One Level decomposition is shown in Fig.2.

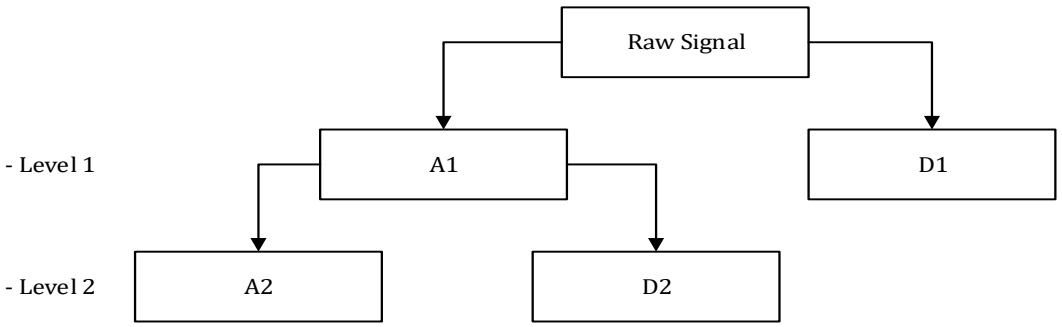

Fig. 2. DWT with two levels of decomposition

The implementation of LPF and HPF shown in Eq. (4) and (5) respectively.

$$A[n] = \sum_{k=-\infty}^{\infty} s[k]g[2n-k] \quad (4)$$

https://doi.org/10.31436/iiumej.v23i1.1760

$$D[n] = \sum_{k=-\infty}^{\infty} s[k]h[2n-k] \quad (5)$$

Then, the signal detail coefficient (D1) was used as a feature vector for further extraction using MFCC. The process of feature extraction using MFCC is shown in Fig. 3.

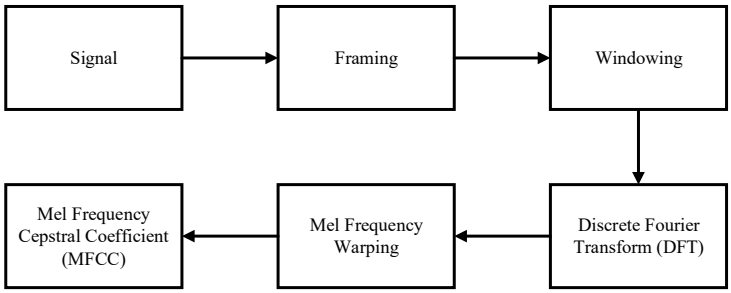

Fig. 3. MFCC feature extraction

One of the most used and popular algorithms for speaker recognition is MFCC, which is also known for representing the information of the speaker's vocal tract [25]. Extracting features using MFCC consists of several steps: framing, windowing, discrete Fourier transform (DFT), Mel frequency warping, and Mel frequency cepstral coefficient.

The framing step, also called signal segmentation, is a process to divide the signal into equal sizes of frames; hence, the voice signal was the non-stationary signal. This process assumed the signal as a stationary signal for a short duration. The frame size used in this research was N = 0.025 seconds, with the overlap M = 0.01 second. Fig. 4 shows how framing works. The framing process caused discontinuity in the framed signal; thus, the windowing function was applied to remove it. We used Hamming window function to eliminate the discontinuity. In terms of calculating the Hamming window, Eq. (6) was used.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) \quad (6)$$

Where $0 \leq n \leq N$ and N is frame length.

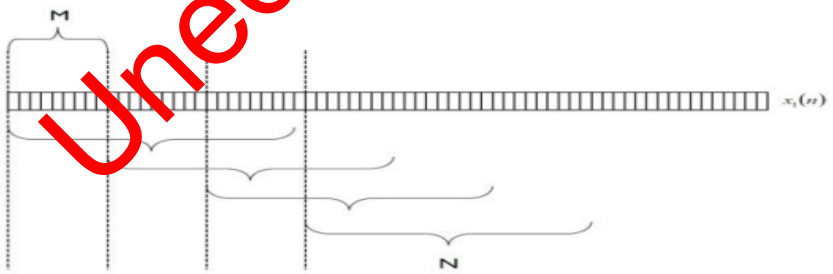

Fig. 4. Framing step

Transforming the signal from the time domain into the frequency domain was the next step. The process was applied using discrete Fourier transformation. We used a total of 512 FFT points. The algorithm is shown in Eq. (7).

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N} \quad (7)$$

Where $X[k]$ is the result of DFT and $x[n]$ is the n-th discrete signal.

After transforming the signal using DFT, the next step was warping the signal using a triangle filter. This process is called Mel frequency warping. This was applied because the signal sound was different from the perception of human hearing, where signal sound did not

https://doi.org/10.31436/iiumej.v23i1.1760

have frequencies on a linear scale. Regarding the Mel frequency value, *mel*(*f*) and triangle filter $H_m[k]$ were used. The following Eq. shows how to calculate *mel*(*f*) and $H_m[K]$, respectively.

$$mel(f) = 1125 \ln\left(1 + \frac{f_{hz}}{700}\right) \tag{8}$$

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \frac{(f[m+1] - k)}{(f\{m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1 \end{cases} \tag{9}$$

We applied a 26-triangle filter, which means we obtained 26 coefficients, but for features, we used the first 13 coefficients as MFCC. The first aim of this research was to determine the best wavelet coefficient as a feature when combined with MFCC. To achieve this, we decomposed the signal into levels 1 and 2 using Haar. The second aim was to determine the best wavelet family when combining with MFCC; hence, we employed all wavelet families, decomposed the signal into one level decomposition, and combined only the detailed coefficient channel with MFCC. Fig. 5 shows the process of feature extraction using the combination of wavelet-MFCC.

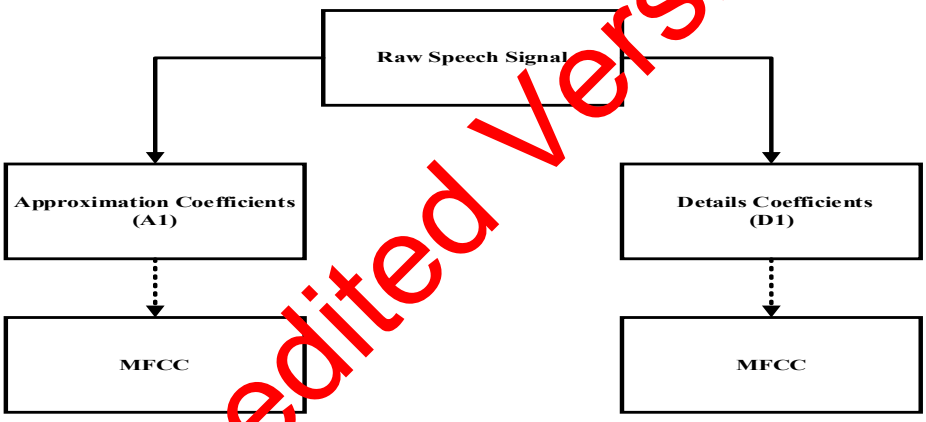

Fig. 5. Wavelet-MFCC feature extraction process.

3.4.Speaker Recognition System

There are various algorithms for speaker recognition, the most used and proven to have excellent performance is Hidden Markov Model (HMM). HMM works by modeling a stochastic process defined by a set of states and several mixtures per state [26]. HMM can solve two characteristics of problems; the first is state-based characteristics, which include the hidden state and the observation state. In the second, there were two types of data consisting of the observation-state sequence and hidden-state sequence [4]. HMM has two types, Ergodic and left-right. The left-right HMM was used to model the speech signal because the speech cannot be repeated to the previous state. Moreover, generally, the observation probability distribution of HMM is modeled by Gaussian Mixture Model (GMM) [19]. The following elements model an HMM:

$$\lambda = (A, B, \pi) \tag{10}$$

A is state transition probability defined as Eq. (11), π is prior probability state distribution defined as Eq. (12), and B is observation probability distribution defined as Eq. (13). This definition represents the left-right type of HMM.

https://doi.org/10.31436/iiumej.v23i1.1760

$$A = \{a_{ij}\} \tag{11}$$

Where, $a_{ij} = 0$ for $j < i$, and $j > i + 1$. $a_{NN} = 1$, $a_{Nj} = 0$, for $j < N$.

$$\pi = \{\pi_i\} \tag{12}$$

Where, $\pi_i = 1$ for $I = 1$, and $\pi_i = 0$ for $I \neq 0$.

$$B = b_j(k) \tag{13}$$

Generally, the HMM observation probability distribution is defined as Gaussian Mixture Model (GMM). However, in practice, this model gives problems in computation. Therefore, the Euclidean distance probability approach can be the alternative solution [19]. It is defined as Eq. (14) and Eq. (15):

$$d(O_t, \mu_j) = \sqrt{\sum_{k=1}^{M} \left(\frac{1}{1+d(O_t,\mu_j)}\right)^2} \tag{14}$$

$$d(O_t, \mu_j) = \sqrt{\sum_{k=1}^{M} (O_{tk} - \mu_{jk})^2} \tag{15}$$

The built model of the recognition system is shown in Fig. 6. The signal was first decomposed into one level using wavelet, and then the sub-band detail coefficient was used as a feature for further extraction using MFCC. The probability of the DWT-MFCC features then calculated probability to be used as a model.

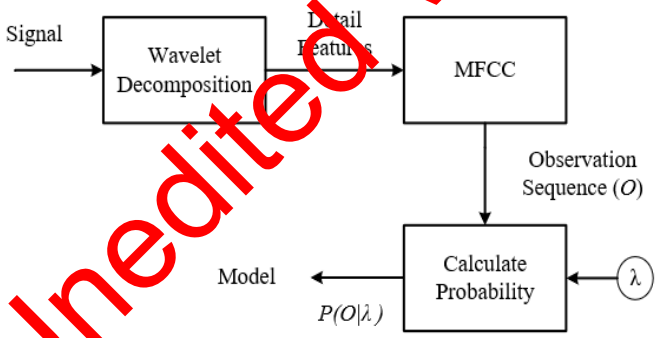

Fig. 6. Speaker recognition system for the training phase.

In this research, the speaker recognition system used is shown in Fig. 7.

https://doi.org/10.31436/iiumej.v23i1.1760

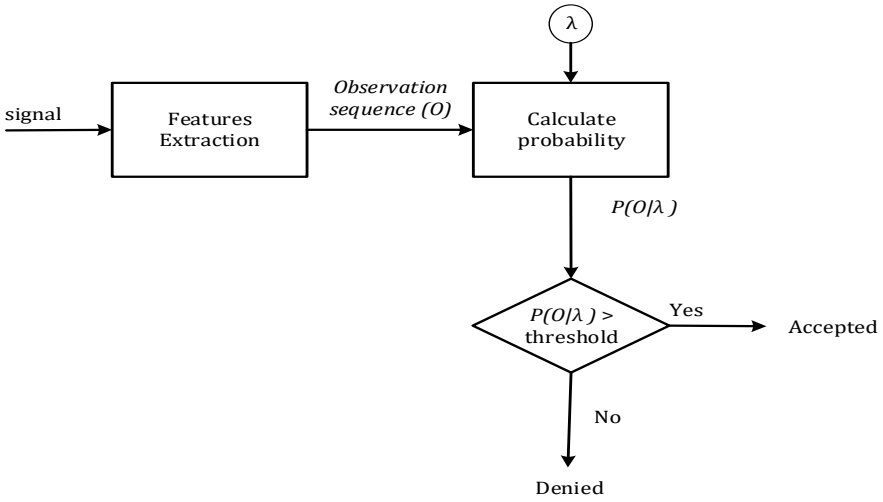

Fig. 7. Speaker recognition system for testing phase

The features of the signal were extracted using the proposed method. The extracted features were then called the observation sequence (O), and fed into the Viterbi algorithm to measure the probability towards lambda (λ). The Viterbi algorithm was employed to find the most likely sequence of hidden states, and the result was an observed sequence. Lambda is a probability sequence obtained from the training model. The model was then compared to the observation sequence to obtain the classification result.

3.5. Evaluation and Validation

The accuracy of the classification result was evaluated using Eq. (16) [2], [27]:

$$Accuracy = \frac{NSRC}{TNS} \times 100 \quad (15)$$

$NSRC$ is the number of speakers recognized correctly and TNS is the total number of speakers.

The K-fold validation method was applied to keep the variation of accuracy low; this method divided the dataset into K datasets. The first dataset was used as a testing dataset, while the second and third K were used as training datasets. Furthermore, the process was repeated, where the second dataset used as testing datasets, the first, third K were used as training datasets. The process was repeated until K times, and the total accuracy is the number of accuracies divided by K. In this research, the number of folds used is 5; thus, in each process, the testing dataset was 20% of the total dataset [28].

4. RESULT AND ANALYSIS

This research proposed a feature extraction algorithm for speaker recognition based on the combination of wavelet-MFCC. We used 300 sound signals obtained from 30 male and female adult volunteers uttering the word “HADIR” in Indonesian, and each volunteer repeated the word ten times. Each signal was then preprocessed to enhance the quality by applying pre-emphasis and normalization.

4.1. Determining the Best Wavelet Coefficient

To obtain the result for the first aim, we employed one- and two-level decomposition using the Wavelet Haar family on both levels of decomposition. The coefficients A1, D1, and a combination of coefficients A1 and D1 were used as features. The next step was extracting those coefficients using MFCC, 13 first coefficients were then used as features. The recognition process was modeled using HMM. We obtained accuracy, as shown in Table 2.

https://doi.org/10.31436/iiumej.v23i1.1760

Table 2: Accuracy Result

Coefficient	Accuracy (%)	
	Level 1	Level 2
A	92	90
D	95	78
A+D	95	85

The highest accuracy was obtained on the one-level decomposition using coefficient D and the combination of A + D, where the accuracy was 95%. Nevertheless, increasing the decomposition level reduced the accuracy of the system recognition, where the highest accuracy obtained was 90%. Hence, it is not necessary to increase the decomposition level as it reduces accuracy. Moreover, the use of the sub-band detail coefficient yielded the best accuracy.

4.2. Determining Best Wavelet Family

Once we figured out the best coefficient, we employed all wavelet families and decomposed the signal into one level decomposition. This allowed us to determine the best wavelet family. The one-level decomposition yielded two sub-bands, approximation and detail. The next step was finding the MFCC of the detail coefficient obtained from the previous step. This way, we obtained wavelet-MFCC features. In order to determine the best family wavelet type when combined with MFCC, we employed all family wavelet types. Then, in the recognition step, we employed HMM to build the model; we evaluated its accuracy using the k-fold cross-validation method described above. The recognition results are shown in Table 3.

Table 3: Recognition result of Wavelet-MFCC for each of Wavelet Family

Type	Acc (%)	Type	Acc (%)	Type	Acc (%)	Type	Acc (%)	Type	Acc (%)
bior 1.1	96.33	coif5	92.33	db11	94.33	db32	92.00	symlet16	93.33
bior 1.3	94.67	coif8	94.33	db12	93.67	db33	93.00	symlet17	93.67
bior 1.5	94.67	coif9	93.00	db13	93.33	db34	91.33	symlet18	93.00
bior 2.2	94.33	coif10	91.67	db14	93.67	db35	91.00	symlet19	92.67
bior 2.4	93.33	coif11	91.33	db15	92.33	db36	91.00	symlet20	94.33
bior 2.6	93.00	coif12	94.00	db16	93.00	db37	93.00	rbio1.1	96.33
bior 2.8	94.33	coif13	90.33	db17	92.67	db38	93.00	rbio1.3	94.67
bior 3.1	94.00	coif14	92.00	db18	92.33	symlet2	95.33	rbio1.5	92.67
bior 3.3	93.67	coif15	93.00	db19	93.00	symlet3	92.67	rbio2.2	95.33
bior 3.5	93.33	coif16	92.00	db20	91.67	symlet4	93.67	rbio2.4	94.00
bior 3.7	94.00	coif17	91.00	db21	91.67	symlet5	92.67	rbio2.6	92.67
bior 3.9	93.67	**db1**(haar)	**96.67**	db22	91.33	symlet6	93.67	rbio2.8	94.00
bior 4.4	93.33	db2	95.33	db23	92.33	symlet7	93.00	rbio3.1	95.33
bior 5.5	93.33	db3	93.67	db24	90.33	symlet8	92.33	rbio3.3	93.33
bior 6.8	94.00	db4	94.33	db25	93.33	symlet9	93.33	rbio3.5	93.00
coif1	95.00	db5	93.67	db26	91.33	symlet10	94.33	rbio3.7	94.67
coif2	93.33	db6	93.67	db27	91.67	symlet11	91.67	rbio3.9	93.00
coif3	92.00	db7	94.67	db28	91.33	symlet12	93.00	rbio4.4	93.67
coif4	92.67	db8	94.00	db29	92.00	symlet13	93.33	rbio5.5	92.67
coif5	95.00	db9	94.33	db30	91.33	symlet14	93.00	rbio6.8	94.00
coif6	93.33	db10	93.33	db31	92.33	symlet15	92.67	dmey	94.33

https://doi.org/10.31436/iiumej.v23i1.1760

exp (added terms): bior – Wavelet Biorthogonal, coif – Wavelet Coiflet, db – Wavelet Daubechies, sym – Wavelet Symlet, rbio – Wavelet Reverse biorthogonal, dmey – Wavelet Dmeyer

4.3. Gender Influence in Speaker Recognition

In this research, we also experimented to find out the influence of the speaker's gender in the recognition system, and there were two focuses on evaluating the influence of gender. First was the influence of gender on each of the wavelet families as a coefficient detail. Second was the influence of gender on each of extracting method, such as MFCC, MFCC + delta, and MFCC + delta-delta.

In the first condition, as shown in Table 4, we obtained better average accuracy for female speakers than the male speakers for each wavelet family. The obtained average accuracy was about 96% for females and about 92% for male speakers. The low accuracy of male speakers was due to the large gap between each wavelet family's highest and lowest accuracy. The obtained gap was up to 8%. While on the female speaker, the gap between the highest and lowest accuracy was up to 5%, where the highest accuracy was 98.17%, and the lowest was 93.00%.

Table 4: Gender influence the wavelet families

Wavelet Family	Average Accuracy (%)					
	Male	Female	Max Male	Min Male	Max Female	Min Female
Daubechies	90.89	95.76	95.61	87.44	98.17	93.83
Coiflet	91.19	95.11	93.50	88.00	98.17	93.17
Symlet	91.49	95.80	93.50	90.22	98.17	93.00
Biorthogonal	92.41	96.41	95.06	90.78	98.17	93.83
Rbiorhogonal	92.24	95.58	95.11	90.22	98.17	94.83
Dmeyer	92.89	96.50	92.89	92.89	96.50	96.50
Average	91.85	96.03	94.28	89.93	97.89	94.19

Nevertheless, from each wavelet family, it can be said that the best accuracy was obtained using Wavelet Daubechies. The obtained result is shown in Table 3, and Table 4 stated that Wavelet Daubechies is the best wavelet family to combine with the proposed method to create Wavelet-MFCC feature extraction.

In the second condition, based on the testing result in Table 5, we obtained the highest average accuracy using the proposed method in male speaker recognition, where the average accuracy was 95.61%. However, for female speaker recognition, there was no difference between the methods; the average accuracy was varied for each of the methods. The accuracy of male and female speakers was influenced by the acoustical characteristic, fundamental frequency or F0, formant frequency, vocal tract length, and vocal fold. The obtained result was in line with the previous research by [29], [30].

Table 5: Gender Influence Accuracy

Method	Accuracy (%)	
	Male	Female
Proposed Method (db1- MFCC)	95.61	98.17
MFCC	94.00	99.00
MFCC + Delta	94.50	99.00
MFCC + Delta-Delta	93.00	98.00

https://doi.org/10.31436/iiumej.v23i1.1760

In the previous research, the recognition of male speakers gives better accuracy than for the female speakers. It was believed that the result was influenced by the F0 of females being higher than males [31]. In males, F0 was around 110 Hz in the range of 100-146 Hz, while in females, around 200 Hz in the range of 188-221 Hz.

4.4. Comparison of other Features Extraction Method

Table 6 shows the comparison of the proposed method compared to conventional MFCC, MFCC + delta, and MFCC + delta-delta.

Table 6: Comparison of Features Extraction Method

Methods	Accuracy (%)
Proposed Method (db1 - MFCC)	96.67
MFCC	95.67
MFCC + Delta	96.00
MFCC + Delta-Delta	94.67

In this research, we proposed using a detail coefficient as a feature in the combination of wavelet-MFCC feature extraction for speaker recognition. The voice that we heard was obtained from the combination of voice sound, resonance, and articulation. The voice sound was the output of the vocal fold vibration; the voice sound was amplified and modified by the vocal tract resonator. The vocal tract resonator includes the throat, mouth cavity, and nasal passage. This resonator makes the sound that distinguishes individual persons. Whereas articulation is the modified process to produce recognizable words.

The resonator can distinguish each person because the produced sound has different variability frequencies. In male and female sounds, the variability lies in high frequency. Specifically, the speaker's identity lies in the area where the formant frequency is higher [32]. The higher frequency is generally analogized as noise, but in the case of sound, the noise component plays an essential role in perceiving sound quality. This noise component is usually modeled as white noise, where the spectrum is not flat and portrays the different shapes of spectral [33]. It is affected by the glottal opening, flow rate, and the shape of the vocal tract. The interaction among spectral shape, the relative level of harmonic, and noise energy in the sound source influence the perception of the quality of the sound. Hence, it can be concluded that the use of wavelet decomposition separates the voice sound and resonance where the voice sound lies in the approximation coefficient, whereas the resonance, which is the speaker identity, lies in the detail coefficient.

5. CONCLUSION

In this research, we employed wavelet-MFCC to extract the features from the sound signal and HMM for recognition. There were several aims in this research. First, determining the best sub-band coefficient. According to the experiment result, the best sub-band was sub-band detail in one level decomposition, where the accuracy was 95%. Once we found out the best sub-band, the second aim was to determine the best wavelet family combined with MFCC. From 105 types of wavelet families, we obtained the best wavelet family, Daubechies order 1 (Haar), where the accuracy was 96.67%. Although the gender type of the speaker also influenced the type of wavelet family, we experimented to find the best wavelet family for recognition of each gender. Based on the result, we concluded that the wavelet Daubechies was the best wavelet family to extract the features on both genders. Therefore, it can be said that from the obtained

result, wavelet Daubechies was the best wavelet coefficient to combine with MFCC. Then, we compared the proposed method to other methods such as conventional MFCC, MFCC + delta, and MFCC + delta-delta. From the comparison, we obtained that the best method for feature extraction was the combination of db1 – MFCC.

For future research, the proposed method needs to be evaluated using more speaker datasets and in noisy environments so the system's robustness can be seen. In addition, the proposed method needs to be investigated in the speaker-independent recognition system, emotion, gender, and speech recognition. Finally, real-time speaker recognition for the proposed method also needs to investigated further.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Research and Technology of the Republic of Indonesia through the Applied Research scheme.

REFERENCES

[1] Tirumala SS, Shahamiri SR, Garhwal AS, Wang R. (2017). Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 90:250-271. doi:10.1016/j.eswa.2017.08.015

[2] Alsulaiman M, Mahmood A, Muhammad G. (2017). Speaker recognition based on Arabic phonemes. Speech Communication, 86:42-51. doi:10.1016/j.specom.2016.11.004

[3] Shaver, Clark D. and Acken, John M. (2016). A Brief Review of Speaker Recognition Technology. Electrical and Computer Engineering Faculty Publications and Presentations. 350. http://pdxscholar.library.pdx.edu/ece_fac/350

[4] Wei, Y. (2020). Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization. IEEE Access, 8: 34942-34948. doi:10.1109/ACCESS.2020.2972511

[5] Huang, Xuedong and Acero, Alex, Hon H.-W. (2001) Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Upper Saddle River, NJ, United States.

[6] Zhao X, Wang Y, Wang D. Robust. (2014). Speaker Identification in Noisy and Reverberant Conditions. ICASSP, IEEE Int Conf Acoust Speech Signal Process – Proc, 22(4):3997-4001. doi:10.1109/ICASSP.2014.6854332

[7] Ayoub B, Jamal K, Arsalane Z. (2016). Gammatone frequency cepstral coefficients for speaker identification over VoIP networks. 2016 Int Conf Inf Technol Organ Dev IT4OD. doi:10.1109/IT4OD.2016.7479293

[8] Daqrouq K, Al Azzawi KY. (2012) Average framing linear prediction coding with wavelet transform for text independent speaker identification system. Comput Electr Eng, 38(6):1467-1479. doi:10.1016/j.compeleceng.2012.04.014

[9] Amelia F, Gunawan D. (2019). DWT-MFCC Method for Speaker Recognition System with Noise. 7th Int Conf Smart Comput Commun ICSCC 2019, pp.1-5. doi:10.1109/ICSCC.2019.8843660

[10] Lung SY. (2007). Efficient text independent speaker recognition with wavelet feature selection based multilayered neural network using supervised learning algorithm. Pattern Recognit, 40(12):3616-3620. doi:10.1016/j.patcog.2007.05.010

[11] Wu J Da, Lin BF. (2009) Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Syst Appl, 36(2 PART 2):3136-3143. doi:10.1016/j.eswa.2008.01.038

[12] Kumar P, Chandra M. (2011). Hybrid of wavelet and MFCC features for speaker verification. Proc 2011 World Congr Inf Commun Technol WICT 2011, pp. 1150-1154. doi:10.1109/WICT.2011.6141410

[13] Turner C, Joseph A. A. (2015). Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification. Procedia Comput Sci, 61:416-421. doi:10.1016/j.procs.2015.09.177

[14] Kishore KVK, Sharrefaunnisa S, Venkatramaphanikumar S. (2015). An efficient text dependent speaker recognition using fusion of MFCC and SBC. 1st Int Conf Futur Trends Comput Anal Knowl Manag ABLAZE 2015, (Ablaze):18-22. doi:10.1109/ABLAZE.2015.7154960
[15] Rathor S, Jadon RS. (2017). Text indpendent speaker recognition using wavelet cepstral coefficient and butter worth filter. 8th Int Conf Comput Commun Netw Technol ICCCNT 2017, pp.1-5. doi:10.1109/ICCCNT.2017.8204079
[16] Badrit N, Tadjt ABC, Gargourt C, Ramachandrant K. (2002). On the use of wavelet and Fourier transforms for speaker verification. The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002., Tulsa, OK, USA, pp. III-344. doi: 10.1109/MWSCAS.2002.1187043
[17] Adam TB, Salam MS, Gunawan TS. (2013). Wavelet based cepstral coefficients for neural network speech recognition. IEEE ICSIPA 2013 - IEEE Int Conf Signal Image Process Appl, pp.447-451. doi:10.1109/ICSIPA.2013.6708048
[18] Rozario MS, Thomas A, Mathew D. (2019). Performance Comparison of Multiple Speech Features for Speaker Recognition using Artifical Neural Network. 9th International Conference on Advances in Computing and Communication (ICACC), Kochi, India, pp. 234-239. doi: 10.1109/ICACC48162.2019.8986182 19.
[19] Hidayat S. Hidayat R. Adji TB.(2016). Speech Recognition of Kv-Patterned Indonesian Syllable Using Mfcc, Wavelet and Hmm. Kursor, 8(2):67. doi:10.28961/kursor.v8i2.63
[20] Sharma G, Umapathy K, Krishnan S. (2020). Trends in audio signal feature extraction methods. Appl Acoust. 158:107020. doi:10.1016/j.apacoust.2019.107020
[21] Sunitha C, Chandra E. (2015). Speaker recognition using MFCC and improved weighted vector quantization algorithm. International Journal of Engineering and Technology, 7(5):1685-1692.
[22] Adam TB, Salam MS, Gunawan TS. (2013). Wavelet Cepstral Coefficients for Isolated Speech Recognition. TELKOMNIKA, 11(5):2731-2738. doi:10.12928/telkomnika.v11i5.2510
[23] Hidayat S, Hasanah U, Rizal AA. (2016). Algoritma Penghapus Derau / Silence Dan Penentuan Endpoint Dengan Nilai Ambang Terbobot Untuk Sinyal Suara. In: Seminar Nasional APTIKOM (SEMNASTIKOM). pp.320-323.
[24] Sekkate S, Khalil M, Adib A. (2018). Using wavelet and short-term features for speaker identification in noisy environment. Int Conf Intell Syst Comput Vision, ISCV 2018. 2018-May:1-8. doi:10.1109/ISACV.2018.8354030
[25] Jahangir, R. Teh, YW. Memon, NA. Mujtaba, G. Zareei, M. Ishtiaq, U. AKhtar, MZ. Ali, I. (2020). Text-independent Speaker Identification through Feature Fusion and Deep Neural Network. IEEE Access, 8:32187-32202. doi:10.1109/ACCESS.2020.2973541
[26] El-henawy IM, Khedr WI, Elkomy OM, Abdalla AMI. (2014). Recognition of phonetic Arabic figures via wavelet based Mel Frequency Cepstrum using HMMs. HBRC J, 10(1):49-54. doi:10.1016/j.hbrcj.2013.09.003
[27] Maurya A, Kumar D, Agarwal RK. (2018). Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach. Procedia Comput Sci, 125:880-887. doi:10.1016/j.procs.2017.12.112
[28] Picard RR, Cook RD. (1984). Cross-Validation of Regression Models. J Am Stat Assoc, 79(387):575-583. doi:10.2307/2288403
[29] Li L, Zheng TF. (2015). Gender-dependent feature extraction for speaker recognition. In IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2015 - Proceedings., pp.509-513. doi:10.1109/ChinaSIP.2015.7230455
[30] Kanervisto A, Vestman V, Sahidullah M, Hautamaki V, Kinnunen T. (2017). Effects of gender information in text-independent and text-dependent speaker verification. ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc., pp.5360-5364. doi:10.1109/ICASSP.2017.7953180
[31] Titze IR. (1989). Physiologic and acoustic differences between male and female voices. J Acoust Soc Am, 85:1699-1707. doi:https://doi.org/10.1121/1.397959
[32] Lee Y, Keating P, Kreiman J. (2019). Acoustic voice variation within and between speakers. J Acoust Soc Am, 146(3):1568-1579. doi:10.1121/1.5125134
[33] Zhang Z. (2016). Mechanics of human voice production and control. J Acoust Soc Am, 140(4):2614-2635. doi:10.1121/1.4964509