International Journal of Technology 11(2) 411-421 (2020)
Received October 2019 / Revised January 2020 / Accepted April 2020

International Journal of Technology

http://ijtech.eng.ui.ac.id

Improving Accuracy of Isolated Word Recognition System by using Syllable Number Characteristics

Risanuri Hidayat[1*], Anggun Winursito[1]

[1]*Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Jl. Grafika No. 2, Yogyakarta 55281, Indonesia*

Abstract. Studies are constantly developing and improving speech recognition systems, especially their accuracy. This study developed an isolated word recognition system by using the syllable number characteristics of speech signals that will be recognized. First, the syllable number of speech signals to be recognized was detected, and then, the detection results were used to call one of the database groups that matched the syllable number characteristics. This method was designed to reduce the error possibility through a matching process between test data features and database features. This study used Mel frequency cepstral coefficients (MFCC) for feature extraction and the K-nearest neighbor (KNN) method for classification. Three versions of the proposed method were designed. The results showed that version three increased the accuracy by 4% compared to the conventional recognition system. Version three had the fastest computational time compared to the other methods. The addition of syllable detection algorithms in version three increased the computational time by only 0.151 s compared to the conventional MFCC method. The data cut length and threshold value for the filter also influenced the speech recognition system accuracy.

Keywords: Isolated word; K-nearest neighbor (KNN); Mel frequency cepstral coefficients (MFCC); Number of syllables; Speech recognition

1. Introduction

Technology plays a crucial role in human daily life and is developing at a rapid rate. One such technology is speech recognition. Speech recognition technology is being widely used in applications such as mobile phones, home security systems, and global positioning systems. Studies have used speech recognition systems to recognize drones (Shi et al., 2018). Studies on speech recognition systems are continually improving the recognition results. Speech recognition systems use several main stages including feature extraction and classification to identify speech patterns (Dahake et al., 2016). The feature extraction process obtains the characteristics of a sound frame, and the classification process chooses a word by analyzing extracted features (Jo et al., 2016). Mel frequency cepstral coefficients (MFCC) are widely used for feature extraction. Some studies have already used MFCC for feature extraction (Adiwijaya et al., 2017; Vijayan et al., 2017; Hidayat et al., 2018; Kumar et al., 2018; Marlina et al., 2018; Winursito et al., 2018; Li et al., 2020). Although MFCC is widely used in speech recognition systems, they still require further development, especially in terms of their accuracy (Winursito et al., 2018).

*Corresponding author's email: risanuri@ugm.ac.id, Tel.: +62-274-552305
doi: 10.14716/ijtech.v11i2.3678

Several studies have tried to improve the performance of the MFCC method. One study improved the MFCC method by adding a delta coefficient (Hossan et al., 2010) and compared this MFCC + Delta method with the ordinary MFCC method. The results indicated that the added delta coefficient improved the speech recognition system's accuracy. Another study (Hidayat et al., 2018) added a wavelet-transform-based noise reduction system. This is because the MFCC is quite susceptible to noise interference in the sound input, and this impacts the speech recognition system's accuracy. Other studies used wavelets and a psychoacoustic model for speech compression (Gunjal and Raut, 2015). A study on noise removal in speech signals (Tomchuk, 2018) tried to realize high speech recognition system accuracy for both signals with and without noise. A recent speech recognition system added a data compression method (Winursito et al., 2018) and compressed the total output data of all MFCC features by using a principal component analysis (PCA) method. Data compression was performed for removing unnecessary data and leaving behind only important data. The study results indicated increased accuracy at the cost of increased computational time.

Many studies have focused on improving the speech recognition system accuracy, generally by adding algorithms from other methods into the system; this resulted in side-effects such as increasingly heavy computational loads. Overly large computational loads are a problem for speech recognition systems because speech recognition applications are expected to work in real-time. The present study aims to improve the speech recognition system accuracy without requiring large computational loads. This study increases the accuracy of an isolated word recognition system by using the syllable number characteristics of the speech to be recognized. Most studies on speech recognition systems use utterance data in the form of isolated words (Masood et al., 2015; Hidayat et al., 2018; Raczynski, 2018; Sawant and Deshpande, 2018; Tomchuk, 2018; Winursito et al., 2018). Others developed syllable-based speech recognition systems (Can and Artuner, 2013; Soe and Theins, 2015; Kristomo et al., 2017). Isolated word recognition systems are preferable because they have high accuracy and require less-complicated algorithms; however, most developed systems add other methods to the system. In this study, the utterance data to be recognized is an isolated word. An isolated word recognition system was developed by using the syllable number characteristics of speech signals to be recognized as additional feature data to increase the speech recognition system accuracy. An added syllable detection algorithm is simplified to avoid greatly affecting the computational load. The word utterance objects examined in this study were several isolated words in Bahasa. In general, daily spoken words are divided into five types of words based on the number of syllables, namely, words that have 1 (one) syllable, 2 (two) syllables, 3 (three) syllables, 4 (four) syllables, and 5 (five) syllables. Researchers use these characteristics to improve the isolated word recognition system accuracy by grouping words into several databases based on the syllable numbers. Three versions of the proposed method were designed to determine the best accuracy improvement. These three versions differed in terms of the division of database group numbers in the classification process. In version one, the database was divided into five parts per syllable number. In version two, the database was divided into three parts. Finally, in version three, the database was divided into only two parts. Test results obtained using the proposed method were then compared with previously developed methods such as MFCC + Delta and MFCC + PCA and in terms of the accuracy and computational time.

2. Methods

The speech recognition process started with recording the speech signals. Then, filtering was performed to eliminate silent signals in the recorded voice signals, leaving behind only the speech information signal. In the conventional method, the filtered signal is cut and used as an input for feature extraction and classification. In the proposed method, the filtered speech signal data were detected by the number of syllables based on the voice signal data length. The syllable detection process produced five types of syllable groups, ranging from one syllable to five syllables. In the proposed method, a syllable number was used for classification to improve the speech recognition system accuracy. After the syllable number was detected, the speech signal data were cut to standardize their size. Feature extraction was then performed using the MFCC method. Next, the extracted feature data were classified using the K-nearest neighbor (KNN) method. In the classification process, the extracted feature data were compared with suitable feature data in the database group based on the syllable detection results obtained from the previous process. With identical syllable number characteristics between test data and reference data in the database group, the error possibility in recognition processing will decrease. Figures 1 and 2 show comparisons of the stages of the speech recognition process between the conventional and the proposed methods.

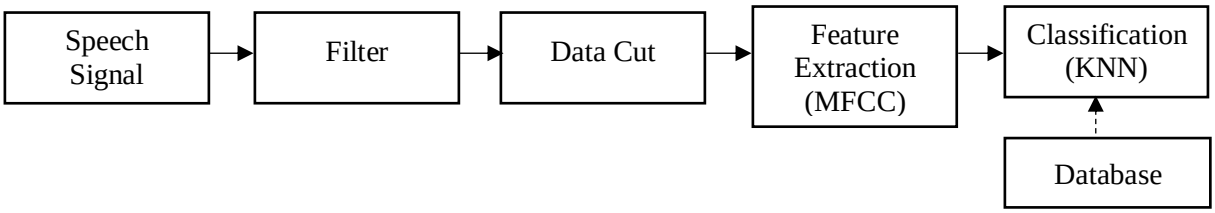

Figure 1 Conventional speech recognition system

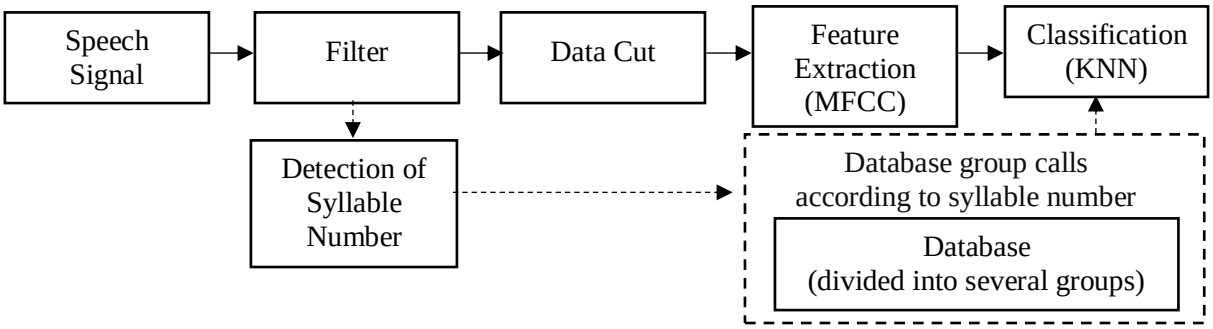

Figure 2 Proposed speech recognition system

In this study, MFCC was used for feature extraction and the KNN method was used for classification. The proposed method uses the syllable number characteristics of each word to be recognized. This study used several isolated words in Indonesian obtained from the Digital Systems Lab, University of Gadjah Mada, that has been used frequently in various studies (Hidayat et al., 2018; Winursito et al., 2018). These data consisted of 25 data classes recorded from seven speakers, for a total to 175 data. All data were then divided into training data and test data. Then, to further test the performance of the proposed method, two test data sets with a total of 50 data were added. Testing was performed by considering the sentence of error (SER) indicator. The data were grouped according to the syllable number. Three versions of the proposed speech recognition system were designed.

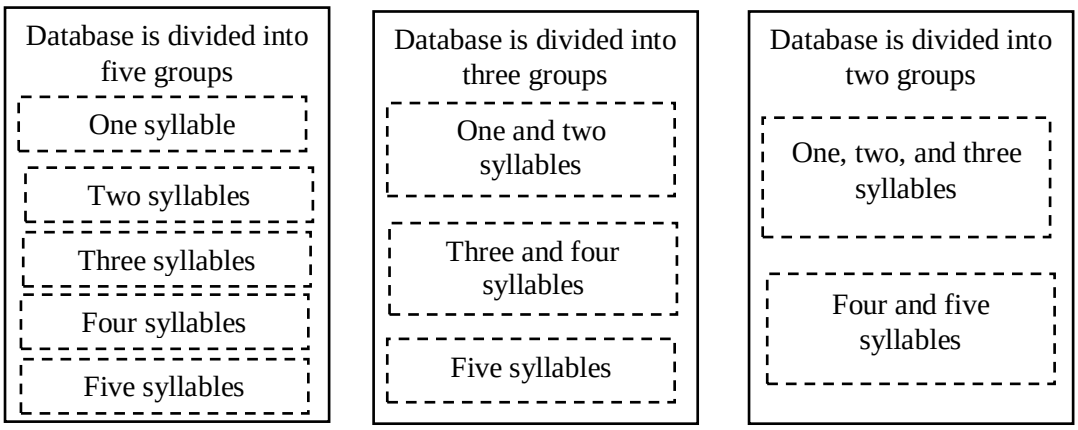

Figure 3 Comparison of database group division in three versions of proposed method: (a) version 1; (b) version 2; and (c) version 3

These three versions differ in terms of the division of database group numbers in the classification process, as shown in Figure 3. In version one, the database was divided into five parts: one, two, three, four, and five syllables. In version two, the database was divided into three parts: one and two syllables, three and four syllables, and five syllables. In version three, the database was divided into only two parts: one, two, and three syllables, and four and five syllables. The database groups were divided based on a syllable detection process. This method was designed to minimize the possibility of errors in the recognition process. By using the syllable number characteristics, a narrower comparison of speech signal features could be made in the classification process, thereby affecting the recognition accuracy.

2.1. Filter

Filters were used to remove silent signals and leave behind only speech information signals. Silent and speech information signals were distinguished by passing signals having a magnitude exceeding a certain threshold. The filter used in this study is an automatic amplitude filter with a threshold of 0.025.

2.2. Syllable Number Detection

The proposed method used the syllable number characteristics of speech signals to be recognized. Therefore, a system for detecting the syllable number of these speech signals was needed. Next, the syllable number characteristics were used to facilitate classification in the next process. Syllable detection was performed by using the speech signal data length. This study calculated 125 data lengths that had been grouped according to the syllable number. Then, the average data length for each syllable group was examined. Furthermore, upper and lower data range limits were determined for each syllable number characteristic by using the average data length values of each syllable group.

2.3. Data Cut

A data cut was performed to homogenize all speech data to be recognized. In this study, several data length values were tested and analyzed to determine the right speech data length by considering the speech recognition system accuracy. The data lengths ranged between 1000 and 10000.

2.4. MFCC Feature Extraction

MFCC is one of the popular feature extraction methods used in speech recognition. It can represent vocal tracts that can be modeled by spectral envelopes in speech segments

(Banaeeyan et al., 2019). The MFCC feature extraction process starts with a pre-emphasis process. The pre-emphasis process used in this study is given by Equation 1.

$$Y[n] = X[n] - (a)X[n-1] \quad (1)$$

where $Y[n]$ is the output signal; $X[n]$, the input signal; and a, the filter coefficient value (generally, 0.95).

The next step after pre-emphasis is a framing process that cuts the signal into small overlapping parts of ~10 ms each to avoid discontinuities in the cut signal (Adiwijaya et al., 2017).

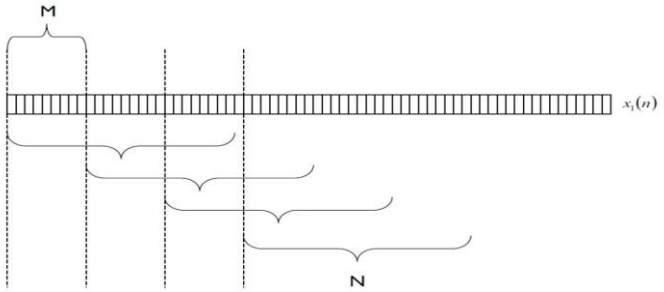

Figure 4 Illustration of the framing process

The framing process in this study used an overlap value of M = 100 and number of sample data per frame value of N = 256, as shown in Figure 4. A Hamming window was used for windowing, as given by Equation 2.

$$(n) = 0.54 - 0.46.\ cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1 \quad (2)$$

Therefore, the output of the windowing process was

$$y1(n) = x1(n)\ w(n),\ \ 0 \le n \le N-1 \quad (3)$$

where $y1(n)$ is the output signal; $x1(n)$, the input signal; $w(n)$, the mathematical equation for the Hamming window; and N, the number of signal samples in each frame. Coefficient values of 0.54 and 0.56 are commonly used in the Hamming window.

The next step in the feature extraction process is a fast Fourier transform (FFT). FFT was performed according to Equation 4.

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \quad (4)$$

where X_n is the FFT output, X_k is the input signal, and N is 0, 1, 2, ..., $N-1$.

After the FFT process, the processed speech data signal was filtered using a Mel filter bank. The Mel filter bank is shaped like a bandpass filter and has linear characteristics below 1000 Hz and logarithmic ones above 1000 Hz (Vijayan et al., 2017).

The Mel frequency scale equation is

$$B\ (f) = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$

where f is the input signal of the result of the FFT process.

The next step in the MFCC feature extraction process is a discrete cosine transform (DCT). Because the Mel coefficient spectrum was a real number, it could be transformed into a time domain by using a DCT.

$$\Sigma_{k=1}^{N}\ log\big(Y(i)\big)\ cos[mx(k-0.5)x\pi \div N] \quad (6)$$

where N is the number of filter bandpass triangles. The m value is between 1 and L, where

L is the number of output coefficients. Y is the input of data cues that are processed using DCT. This study uses 13 MFCC coefficients.

2.5. KNN Classification

The KNN method was used for classification and speech data recognition against reference data. KNN uses a set of data obtained from feature extraction and a target class that will be compared with the score of the data features (Mufarroha and Utaminingrum, 2017).

The Euclidean distance is usually used in the KNN classifier to calculate the similarity between training and test data (Enriko et al., 2016). This study used the Euclidean distance method given by Equation 7.

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \tag{7}$$

where D is the result of the Euclidean distance magnitude, and a and b are two features whose distance is measured.

To evaluate the performance of the proposed method through a comparison with other methods, the sentence error rate (SER) is used. The SER is calculated using equation (8) to determine the performance of each method.

$$SER = \frac{Number\ of\ incorret\ utterances}{Total\ utterances} \tag{8}$$

3. Results and Discussion

3.1. Syllable Number Detection

The proposed method used the syllable number of words to be recognized. Previously, the syllable number detection process was applied to the speech data to be recognized. The syllable number detection system used the data length of each word based on its syllable number. The proposed method was designed in three versions. Before testing the test data, the accuracy of each version was tested for syllable number detection. Table 1 shows the average time characteristics and data numbers of several words based on the syllable number used for determining the parameters of the syllable number detection system.

Table 1 Word characteristics based on syllable number

Syllable number	Average time (s)	Average data length
1	0.37225	2978
2	0.54975	4398
3	0.78462	6277
4	1.051875	8415
5	1.294125	10353

Table 2 shows the characteristics of the data range in each word group based on the syllable number. Further, the accuracy of the three versions of the syllable number detection system is shown. The results showed that version one had the lowest accuracy of 86%. Version two had an accuracy of 92%. Version three had the highest accuracy of 100%, indicating that the syllable number of the test data is recognized correctly.

Table 2 Range limit of data length of three versions of proposed method

Segment	Range limit of data length in proposed method		
	Version 1	Version 2	Version 3
I	<3688	<5109	<7446
II	3689–5108	5109–9384	>7445
III	5109–7445	>9384	
IV	7446–9384		
V	>9384		
Tested accuracy of syllable number detection system	86%	92%	100%

3.2. Data Cut

The next process involved determining the data cut length. The number of data used in the speech recognition process greatly affects the system accuracy; therefore, it was necessary to process the exact number of data cuts.

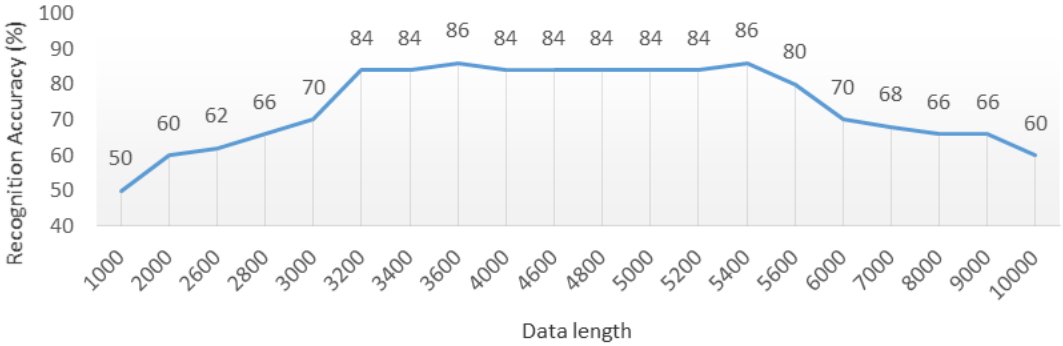

Figure 5 Effects of data length on speech recognition system accuracy

Figure 5 shows the accuracy results for data cut lengths of 1000 to 10000. The results showed that data cut lengths of 3600 and 5400 gave the highest accuracy of 86%. A data cut length of 3600 was selected instead of one of 5400 because a smaller amount of data would likely reduce the computational load.

Table 3 Effect of *k* value on accuracy of KNN classifier in recognition system

k value	Accuracy (%)
1	86
2	70
3	68
4	68
5	72
6	68

The KNN method was used for classification. The determination of the *k* value used in the KNN classifier is very important because it could affect the system accuracy. Table 3 shows some *k* values tested with the system. The results show that a *k* value of 1 provides the highest accuracy.

3.3. Proposed Method

The proposed method uses the syllable numbers of isolated words to be recognized. Three versions are designed to determine the maximum performance. Each version differs in terms of the number of database groups used based on the syllable number characteristics. Each method was tested with 50 test data (test data I). The test data is not used for training. After testing, the accuracy results were compared with those of the recognition system with the conventional MFCC method and the methods that had been developed in previous studies. Table 4 shows a comparison of the performance of the three versions of the proposed method.

Table 4 Comparison of recognition system accuracy of three versions of proposed method

Proposed method	Accuracy (%)
Version 1	82
Version 2	86
Version 3	90

The results show that version three provides the best performance in that the accuracy of syllable number detection is the highest. Version one shows the lowest accuracy of 82%, and version two shows 86% accuracy. Version three shows the best accuracy of 100%. Versions one and two have lower syllable number detection accuracy; this degrades the speech recognition accuracy. In version three, the database is divided into two groups: one, two, and three syllables, and four and five syllables. Furthermore, the best version of the proposed method is compared with other methods, as shown in Figure 6.

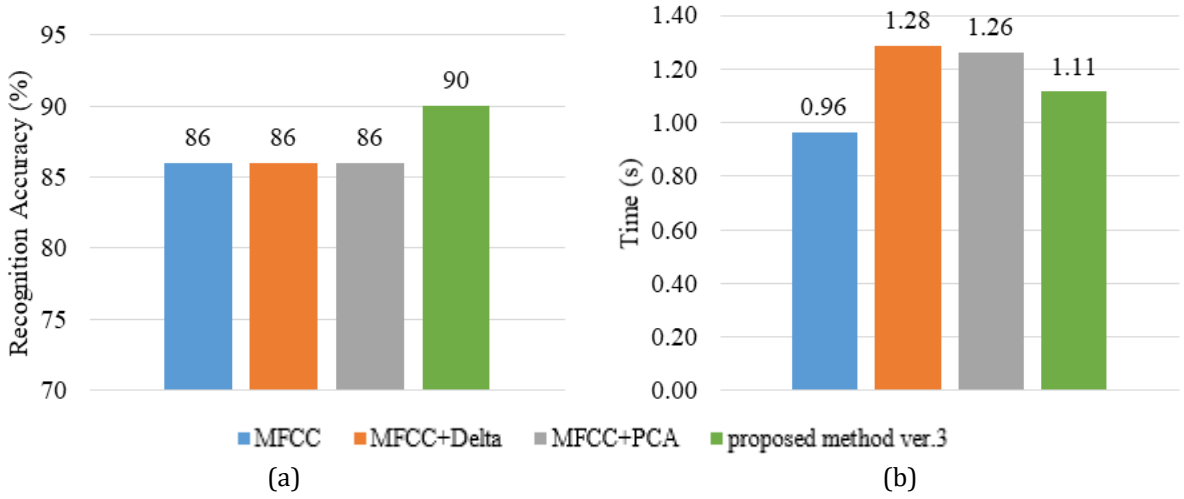

Figure 6 Comparison of conventional and proposed methods in terms of: (a) recognition accuracy; and (b) computational time

Figure 6a shows a comparison of the recognition system accuracy results of several methods: conventional MFCC, MFCC+Delta, MFCC+PCA, and proposed methods. These results show that the proposed method has an accuracy of 90% compared to 86% for the conventional MFCC method. The proposed method shows improved accuracy because it uses certain databases that are adjusted to the characteristics of the voice data to be recognized; the characteristic used is the syllable number. The sound to be recognized first is detected by the syllable number; then, the system calls a specific database according to the number of syllables recognized. Through this mechanism, the proposed method can

minimize recognition errors and increase the accuracy. Figure 6b shows a comparison of the computational times for speech recognition for all methods. The conventional MFCC method has the lowest computational time because it has the simplest algorithm compared to the other methods. Version three, which has the highest accuracy, requires 0.151 s more than the conventional method for computing. This is because the added syllable number detection algorithms increase the computational time. However, compared with other development methods, the proposed method has the fastest computational time. To ensure consistency in the performance of the proposed method, tests are performed using other test data groups. SER is used to determine the performance of all methods. Table 5 shows the test results of several methods; the proposed method achieved the lowest error in all tests. The lower the error rate obtained, the better is the performance of the tested method.

Table 5 SER performance of different feature extraction method using other datasets

Method	SER	
	Test data II	Test data III
MFCC	0.16	0.12
MFCC+Delta	0.12	0.12
MFCC+PCA	0.16	0.12
Proposed method	0.08	0.04

4. Conclusions

The development of speech recognition systems by using syllable number characteristics improved the speech recognition accuracy. Version three of the proposed method improved the speech recognition accuracy by 4% compared to the conventional MFCC method. This method was developed by dividing the reference database into two parts based on the syllable number characteristics. In developing a recognition system using the proposed method, the speech recognition system accuracy strongly depends on the syllable number detection accuracy. That is because if the system incorrectly recognizes the syllable number, the classification process will use a wrong database and the word recognition will also be wrong. The data cut length and threshold values also affected the speech recognition system accuracy. Version three had the fastest computational time compared to other methods. The addition of syllable detection algorithms to version three of the proposed method only increased the computation time by 0.151 s compared with the conventional MFCC method.

References

Adiwijaya, A., Aulia, M.N., Mubarok, M.S., Novia, W.U., Nhita, F., 2017. A Comparative Study of MFCC-KNN and LPC-KNN for Hijaiyyah Letters Pronunciation Classification System. *In*: IEEE Fifth International Conference on Information and Communication Technology (ICoICT), pp. 1–5

Banaeeyan, R., Karim, H.A., Lye, H., Fauzi, M.F.A., Mansor, S., See, J. 2019. Acoustic Pornography Recognition using Fused Pitch and Mel-Frequency Cepstrum Coefficients. *International Journal of Technology*, Volume 10(7), pp. 1335–1343

Can, B., Artuner, H., 2013. A Syllable-based Turkish Speech Recognition System by using Time Delay Neural Networks (TDNNs). *In*: International Conference on Soft Computing and Pattern Recognition (SoCPaR). Hanoi, Vietnam, pp. 219–224

Dahake, P.P., Shaw, K., Malathi, P., 2016. Speaker Dependent Speech Emotion Recognition Using MFCC and Support Vector Machine. *In*: IEEE International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 1080–1084

Gunjal, S., Raut, R., 2015. Traditional Psychoacoustic Model and Daubechies Wavelets for Enhanced Speech Coder Performance. *International Journal of Technology*, Volume 6(2), pp. 190–197

Hidayat, R., Bejo, A., Sumaryono, S., Winursito, A., 2018. Denoising Speech for MFCC Feature Extraction using Wavelet Transformation in Speech Recognition System. *In*: IEEE 10th International Conference on Information Technology and Electrical Engineering (ICITEE), Kuta, pp. 280–284

Hossan, Md.A., Memon, S., Gregory, M.A., 2010. A Novel Approach for MFCC Feature Extraction. *In*: IEEE 4th International Conference on Signal Processing and Communication Systems, Gold Coast, Australia, pp. 1–5

Enriko, I.K.A., Suryanegara, M., Gunawan, D., 2016. Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters. *Journal of Telecommunication, Electronic and Computer Engineering,* Volume 8(12), pp. 59–65

Jo, J., Yoo, H., Park, I.-C., 2016. Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems. *In:* IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Volume 24(2), pp. 754–758

Li, Q., Yang, Y., Lan, T., Zhu, H., Wei, Q., Qiao, F., Liu, X., Yang, H., 2020. MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method with Mixed-signal Processing Architecture for Wearable Speech Recognition Applications. *In:* IEEE Access, Volume 8

Kristomo, D., Hidayat, R., Soesanti, I., 2017. Classification of the Syllables Sound using Wavelet, Renyi Entropy and AR-PSD features. *In*: IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, pp. 94–99

Kumar, C., ur Rehman, F., Kumar, S., Mehmood, A., Shabir, G., 2018. Analysis of MFCC and BFCC in a speaker identification system. *In*: International Conference on Computing, Mathematics and Engineering Technologies (ICoMET), Sukkur, pp. 1–5

Marlina, L., Wardoyo, C., Sanjaya, W.S.M., Anggraeni, D., Dewi, S.F., Roziqin, A., Maryanti, S., 2018. Makhraj Recognition of Hijaiyah Letter for Children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) Method. *In*: IEEE International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, pp. 935–940

Masood, S., Mehta, M., Namrata, Rizvi, D.R., 2015. Isolated Word Recognition using Neural Network. *In*: IEEE Annual IEEE India Conference (INDICON), New Delhi, India, pp. 1–5

Mufarroha, F.A., Utaminingrum, F., 2017. Hand Gesture Recognition using Adaptive Network Based Fuzzy Inference System and K-Nearest Neighbor. *International Journal of Technology*, Volume 8(3), pp. 559–567

Raczynski, M., 2018. Speech Processing Algorithm for Isolated Words Recognition. *In*: IEEE International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, pp. 27–31

Sawant, S., Deshpande, M., 2018. Isolated Spoken Marathi Words Recognition using HMM. *In*: IEEE 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp. 1–4

Shi, L., Ahmad, I., He, Y., Chang, K., 2018. Hidden Markov Model Based Drone Sound Recognition using MFCC Technique in Practical Noisy Environments. *Journal of Communication and Network*, Volume 20, pp. 509–518

Soe, W., Theins, Y., 2015. Syllable-based Myanmar Language Model for Speech Recognition. *In*: IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), Las Vegas, NV, USA, pp. 291–296

Tomchuk, K.K., 2018. Spectral Masking in MFCC Calculation for Noisy Speech. *In*: IEEE Wave Electronics and Its Application in Information and Telecommunication Systems (WECONF), St. Petersburg, pp. 1–4

Vijayan, A., Mathai, B.M., Valsalan, K., Johnson, R.R., Mathew, L.R., Gopakumar, K., 2017. Throat Microphone Speech Recognition using MFCC. *In*: IEEE International Conference on Networks & Advances in Computational Technologies (NetACT), pp. 392–395

Winursito, A., Hidayat, R., Bejo, A., Utomo, M.N.Y., 2018. Feature Data Reduction of MFCC using PCA and SVD in Speech Recognition System. *In*: IEEE International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, pp. 1–6