ISSN: 1978-0524

Speech classification using combination virtual center of gravity and k-means clustering based on audio feature extraction

Diah Kumalasari[a,1] , Arief Bramanto Wicaksono Putra[a,2, *], Achmad Fanany Onnilita Gaffar [a,3]

[a] *Department of Information Technology, Politeknik Negeri Samarinda, East Kalimantan, Indonesia*
[1] *diahkumala06@gmail.com;* [2] *ariefbram@gmail.com;* [3] *onnygaffar212@gmail.com*
** corresponding author*

ABSTRACT

Voice recognition can be done in a variety of ways. Sound patterns can be recognized by performing sound feature extraction. The trainer sound data is built from the best sound data selection using a correlation coefficient based on the level of similarity between sound data for optimal sound features. Extraction of voting features on this research using the Virtual Center of Gravity method. This method calculates the distance between the sound data against the center point of gravity with visualizations in the 3-dimensional form of white, black, and grey pattern spaces. The preprocessing process generates a complex number of data consisting of real numbers and imaginary numbers. The number will be calculated the distance to the Virtual Center of Gravity's pattern space using Euclidean Distance. The sound feature testing is done using K-Means Clustering by means of a speech classification data based sound. The results showed an accuracy of 92.5%.

Keywords:
Classification
Feature Extraction
K-Mean
Virtual Center of Gravity

I. Introduction

The human voice is one of the biometric forms that can be used to recognize a person's character. The process of automatically recognizing spoken words of a speaker based on information in speech signal is called Speech Recognition[1]. Speech recognition is the machine or program's ability to identify words and phrases from spoken language and convert them into machine-readable format[2]. Feature extraction is most important part of the speech recognition system which distinguishes one speech from another[2]. The method of extracting sound features used is the VCG (Virtual Center of Gravity) method.

The Virtual Center of Gravity method uses the concept of Physics Science center of gravity. Center of gravity is an object's heavy distribution center when the center of gravity can be regarded as a style, this is the point where the object is in perfect balanced state no matter how the object is rotated or flipped at the reversed point at that point [3]. This concept is applied in order to find a special feature in an object. This concept is applied in order to find a special feature in an object.

The sound that is captured by the sound recording tool, will mealui several stages of sound processing to obtain the sound feature. Sound feature extraction is the process of converting voice signals into several parameters, where some sound data is considered useless (noise) will be discarded without removing the true meaning of the sound signal [4]. The process used in this study to eliminate sound data that is considered insignificant (noise) is Truncation, normalization, Frame Blocking, Windowing, Fast Fourier Transform (FFT). In this case, the process is done so that the sound data is good enough to be used in the extraction of sound features.

This system will classify the sound data feature into clusters based on the same word pronunciation using K-Means Clustering algorithm [5]. The result of this research system will recognize the voice speech of a person based on the cluster that has formed. The sound identification process is indispensable for knowing the voice greeting's Accuracy based on its features.

This research aims to construct a prototype sound feature using Virtual Center of Gravity in 3-dimensional form and perform a sound feature test to recognize the system using the K-Means clustering process accurately.

DOI: 10.26555/jifo.v14i2.a17390

II. Method

Voice sampling is done by recording sound using the voice recorder app that is on your phone or computer. The voting process is done up to several times.

The system can be divided into audio data collection into two stages: data training and data testing. Data training includes preprocessing I, best sampling, preprocessing II, and audio feature extraction using virtual center of gravity. Data testing includes preprocessing and audio feature extraction using the virtual center of gravity. The system block diagram is shown in Fig. 1.

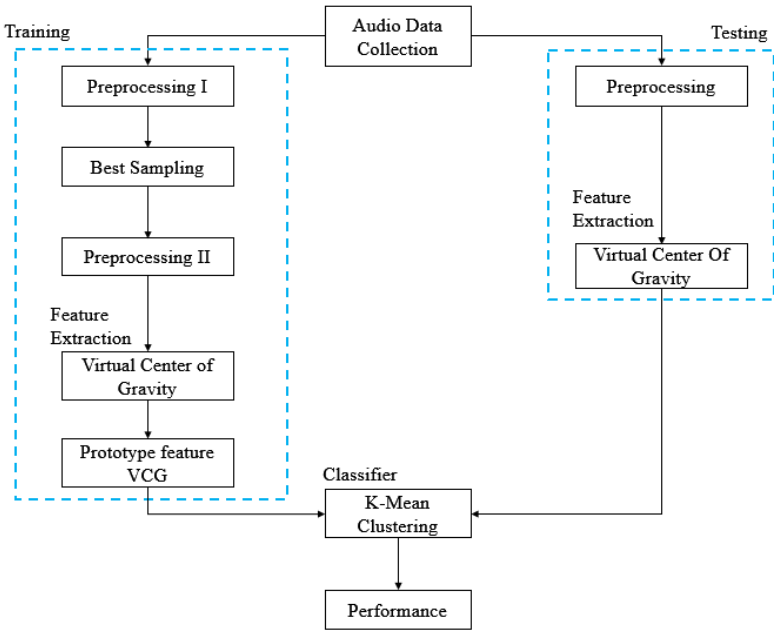


Fig. 1. Block diagram

A. *Preprocessing Level I*

Preprocessing is divided into 2, namely Level I and Level II. On the level I of preprocessing consists of Data Read, Truncation, Data Sampling, and Normalization. Flow preprocessing level I shown in Fig. 2.

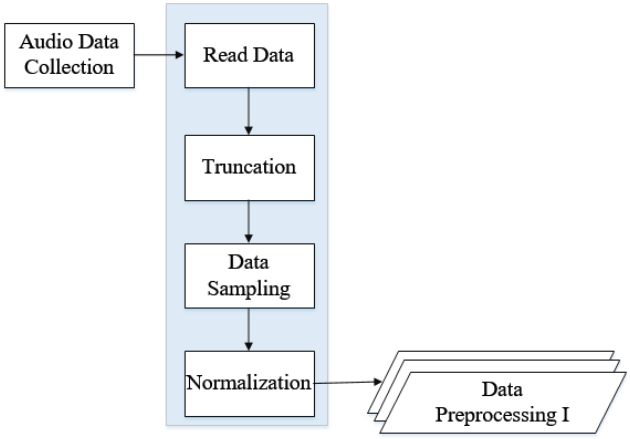


Fig. 2. Flow Preprocessing I

1) Read Data

Voice data reading is performed by the Audioread function located in Matlab. The audio data format used in this study is M4A.

2) Truncation

The sound data truncation is performed to bypass the sound data portion that is not considered necessary. Cutting is done at the beginning and end of the sound signal data, because at the time of

sound recording there is usually a pause that causes the amplitude close to the value 0. To get a voice that only the voice is done, the sound data cuts that have a value of amplitude below 0.01.

3) Data Sampling

Sound data sampling is taking data length data on certain vectors. The sampling is done to homogenate the voice signal data length with the other.

4) Normalization

Feature normalization techniques represent a vital part of each biometric recognition system[6]. Normalization is used to keep the value range amplitude the sound signal does not differ considerably from other voice signals.

$$\bar{x}(n) = \frac{x(n)}{\max(x)} \quad (1)$$

Where $\bar{x}(n)$ denotes result vector normalization of sound signal samples to n. And $x(n)$ denotes n vector audio samples.

B. *Best Sampling*

The best data selection stage aims to get the best feature by optimizing the training data selection. The trainer data is selected through the selection of similarities between other sound sample data. The best data selection flow is shown in Fig. 3.

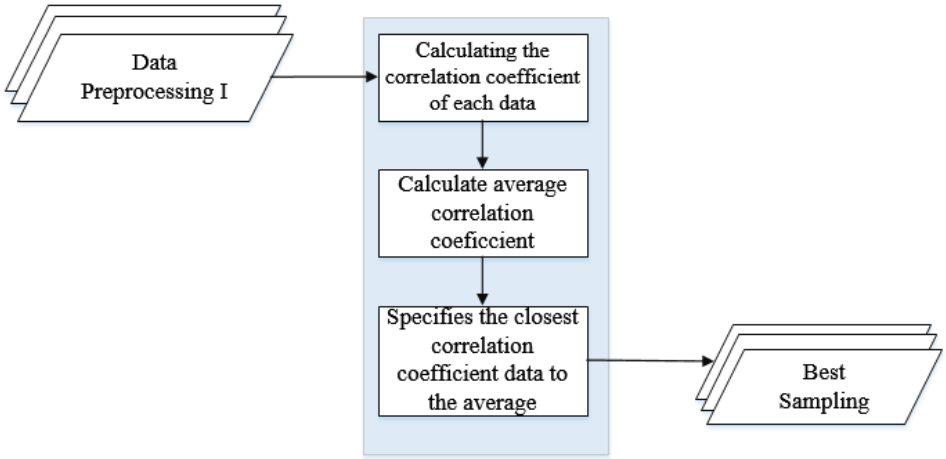


Fig. 3. Flow Best Sampling

The process of selecting this data uses the value of the correlation coefficient. The correlation coefficient is a value that indicates the level of similarity between 2 variables. The value of a correlation coefficient approaching a value of 1 means indicating a very similar level of resemblance. Conversely if the value of a correlation coefficient approaching a value of 0 means that the level of similarity between 2 variables is very low[7].

The calculation of the correlation coefficient applies to all relationships between 2 different sound data variables formed. Then calculate the average of all values of the correlation coefficient. Specifies the value of the corelation coefficient closest to the average value of the correlation coefficient. The level of resemblance between 2 sound data variables, e.g. x and y expressed in the following equation:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-y)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (2)$$

where x_i and y_i are data set x and y, $\bar{x}$ and $\bar{y}$ are average dataset x and y, $\rho_{x,y}$ = -1 where x and y have a perfect negative correlation, and $\rho_{x,y} = 1$ is supposed to x and y have a perfect positive correlation

C. *Preprocessing Level II*

Preprocessing level II consists of Frame blocking, Windowing, and Fast Fourier Transform. Flow preprocessing level I shown in Fig. 4.

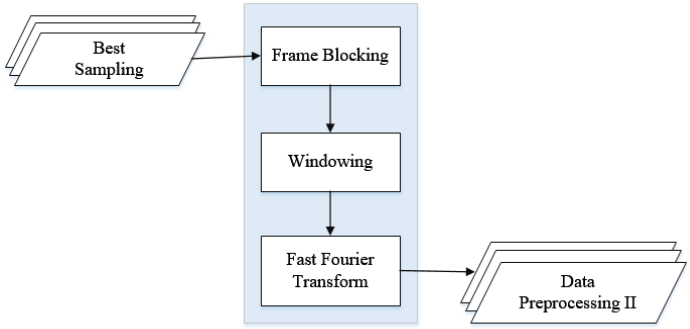


Fig. 4. Flow Preprocessing II

1) Frame Blocking

The sound signal is divided into several frames, where each frame will consist of a sample of the same data. Frame Blocking is generally done overlapping for every frame[8]. Overlapping is done to avoid loss of characteristic or sound characteristics at the border of each frame. The length of overlapping areas, in general, is used is approximately 30% to 50%.

Suppose M is the number of samples between adjacent frames. N is the number of data samples per frame, then M < N. Illustration of the frame blocking shown in Fig. 5.

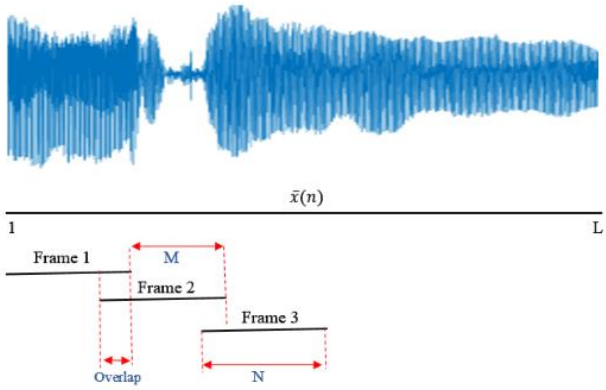


Fig. 5. Illustration Frame Blocking

The overlap is usually expressed in percentages as follows :

$$overlap = \frac{N - M}{N} \times 100\ \%$$

If L is the normalized amount of sampled sound data and W the number of frames, then:

$$L = N + (M).(W - 1)$$

2) Windowing

In the next process is the Windowing. The windowing technique is an important part of the data processing used to find the window's optimum length for the feature extraction process [9]. There are many types of window, e.g., Rectangular, Bartlet, Welch, Hanning, Hamming[10]. The type of Windowing used in this research is the Hamming window. The windowing process reduces unsustainable signals at the beginning and at the end of each frame. Signal generated from the process windowing, expressed in the form of the following equation:

$$y(n) = x(n) \times w(n) \quad (5)$$

Where w(n) uses the Hamming window function, so that the equation becomes:

$$w(n) = 0.54 - 0.46 \times cos\left(\frac{2\pi n}{N-1}\right) \quad (6)$$

3) Fast Fourier Transform

Fast Fourier Transform used for transforming the discrete-time signal from the time domain into its frequency[11] using a formula like the following:

$$\bar{X}_n = \sum_{n=0}^{N-1}(e^{-j2\pi kn/N}).\bar{x}_n \quad (7)$$

These signals essentially represent a signal decomposition in regards to sinusoidal components. Sinusoidal is a sinusoid of the same frequency, but the amplitude and the different phases[12]. FFT is an algorithm developed by CooLey, and Turkey is a signal from the realm of time to be a frequency.

The result of this stage is a complex number consisting of imaginary numbers and real numbers. The number will be used to characterize the pattern space of the sound feature extraction pattern.

D. Audio Feature Extraction

The preprocessing sound Data will be performed extraction feature by using the Virtual Center of Gravity (VCG). This method determines the sound characteristic by looking for the center point of gravity of a pattern space visualized in a 3-dimensional form with black, white, grey, each space of the patterns. Black means the maximum value of an object; white means the minimum value of the object, and gray means the value between the maximum and minimum of the object. The audio feature extraction process flow is shown in Fig. 6.

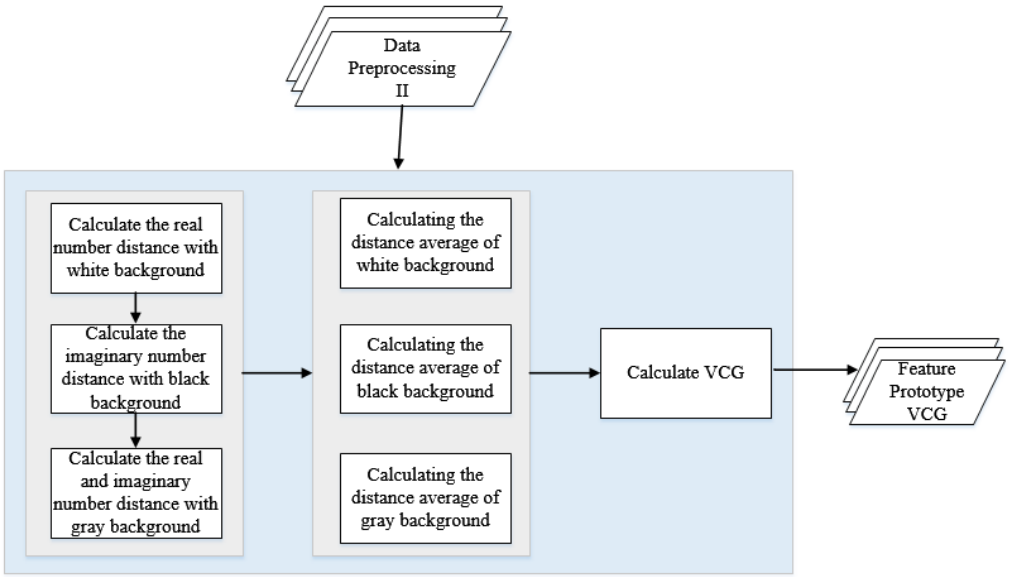


Fig. 6. Flow Feature Extraction

The VCG is a feature representation of the IP Center's of gravity of the feature space (FS/space pattern) and Background (pattern background)[13], the concept of the VCG conducted in this study is explained through the representations shown in Fig. 7.

Representations of real numbers and imaginary numbers that have been formed against virtual pattern spaces are visualized in 3-dimensional form. Virtual Center of Gravity is derived from calculating the distance of real numbers against a white pattern space, an imaginary distance to a black-patterned pattern, and a real and imaginary distance from the gray pattern space. The calculation of this distance uses Euclidean distance.

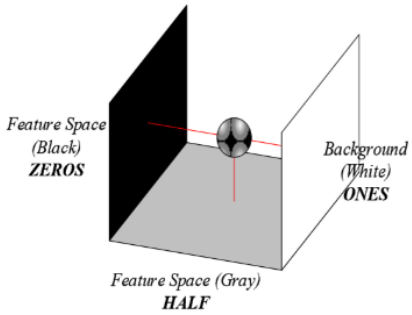


Fig. 7. Virtual Center of Gravity Representation

E. Classifier

The method used for the classification of sound feature is K-Means Clustering. K-Means is one of the algorithms in data mining that can be used to group/clustering data[14]. The K-Means method is a method included in the distance-based clustering algorithm that divides the data into a number of clusters, and this algorithm only works on numeric attributes[15]. The K-Means algorithm is an algorithm that is often used in grouping techniques because it creates an efficient estimate and does not require many parameters. Flow clustering of sound features shown in Fig. 8.

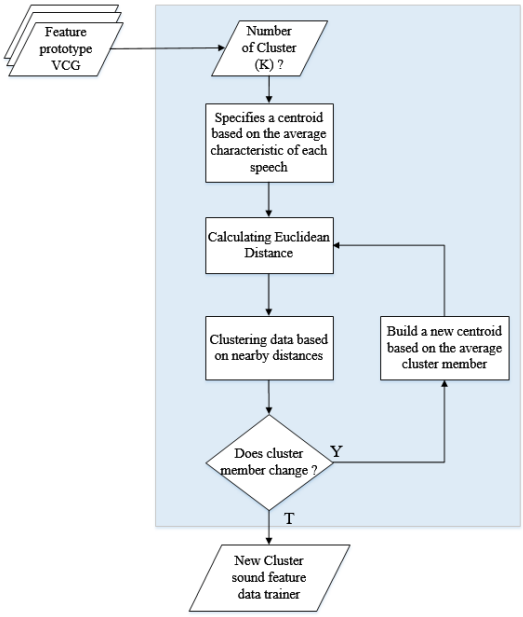


Fig. 8. Flow K-Means Clustering

F. Performance

Sound data that has formed a new cluster will be tested using sound feature test data extraction. The test data is to be P as valid sound data and N as the forgery sound data. The measurement success rate is seen with 2 fault models, namely the False Acceptance Rate (FAR) ratio and False Rejectance Rate (FRR)[16]. Then it is necessary to find the True Positive Rate (TPR), False Positive Rate (FPR), and True Negative Rate (TNR), which are described as follows:

- TPR also called with sensitivity, or accuracy ratio, described as valid match audio hereinafter called True Positive (TP) divided the number of valid audio (P):

$$TPR = \frac{TP}{P} \quad (8)$$

- FPR can also be called alarm error or ratio imprecision, outlined into a valid Unmatch audio hereinafter called False Positive (FP) divided the number of audio forgery (N)

$$FPR = \frac{FP}{N} \quad (9)$$

- TNR can also be called by specificity, described as match forgery audio hereinafter called True Negative (TN) divided the number of audio forgery (N).

$$TNR = \frac{TN}{N} \gg TNR = 1 - FPR \quad (10)$$

- False Acceptance Rate is the value of the False Positive Rate, expressed with the following equation:

$$FAR = FPR \quad (11)$$

- False Rejectance Rate is the value of False Negative Rate, the similarities are:

$$FRR = 1 - TPR \quad (12)$$

- Accuracy (ACC) is a percentage of the Accuracy of the total success submission to the prototype of the stated characteristics with the following equation :

$$Acc = \frac{(TP+TN)}{(P+N)} \times 100\ \% \quad (13)$$

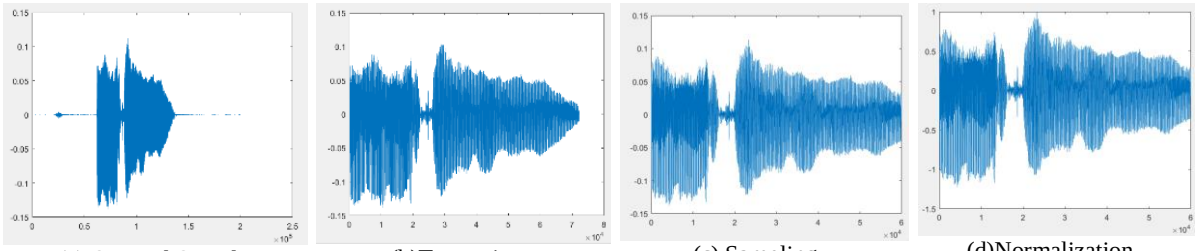
III. Results and Discussion

The sound data of the trainer is obtained from 1 person who speaks the word "Morning", "Certification of Appreciation", "Information technology", each consisting of 10 sound data. As for the test sound data is obtained from 3 respondents who say the word "Morning", "Certification of Appreciation", "information technology" each of 3 sound data. The received sound Data will be created in one folder with the naming of consecutive files shown in Table 1.

Table 1. Sound Data

No	Spoken word	Amount of Data	File Name
1	Morning	10	morning1.m4a s/d morning10.m4a
2	Certification of Appreciation	10	certi1.m4a s/d certi10.m4a
3	Information Technology	10	techno1.m4a s/d techno10.m4a

In the next process of preprocessing level 1 consists of truncation, sampling data, normalization data. The sample result preprocessing level 1 is shown in Fig 9. Then best sampling from each spoken word is shown in Table 2.



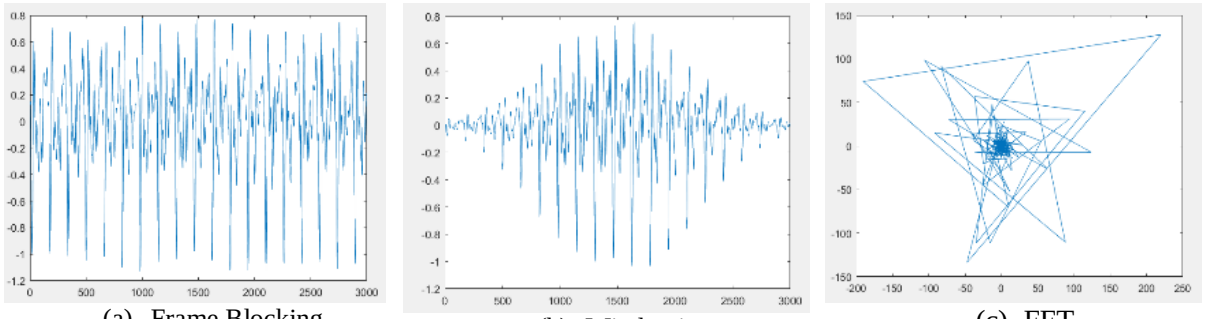
(a) Original Signal (b)Truncation (c) Sampling (d)Normalization

Fig. 9. Sample result preprocessing level I

Table 2. Best Sampling

No	Spoken word	Best Sampling	Sound File Name
1	Morning	1, 2, 3	morning1.m4a, morning2.m4a, morning3.m4a
2	Certification of Appreciation	5, 7, 8	certi5.m4a, certti7.m4a, certi8.m4a
3	Information Technology	3, 5, 9	techno1.m4a, techno5.m4a, techno9.m4a

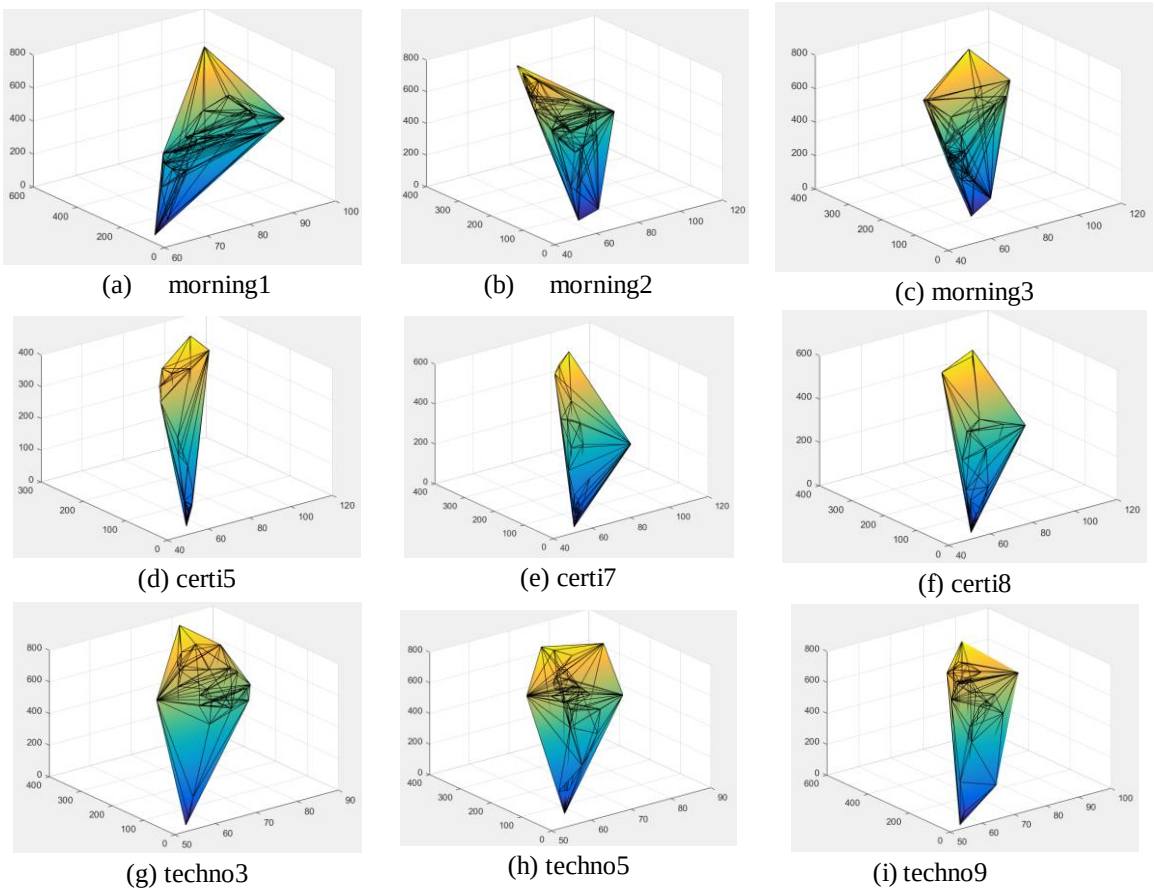
After getting the best training, data will be done preprocessing level 2, namely Frame blocking, Windowing, and Fast Fourier Transform. The sample result preprocessing level 2 is shown in Fig. 10.



(a) Frame Blocking (b) Windowing (c) FFT

Fig. 10. Sample result preprocessing level II

The extraction of sound features using the virtual Center of Gravity is to build a virtual background/pattern space with a 3-dimensional shape that is visualized enameled white, gray, black. The white pattern space value is depicted with a value of 0, the black pattern space value is 1, and the space value of the gray pattern is 0.5. Sound feature extraction generates three sound pattern features of each data. The prototype visualization of a white pattern room sound feature is depicted in blue, the grey pattern space is depicted with a green color, and a black pattern space with the yellow color shown in Fig. 11.



(a) morning1 (b) morning2 (c) morning3

(d) certi5 (e) certi7 (f) certi8

(g) techno3 (h) techno5 (i) techno9

Fig. 11 Prototype Feature

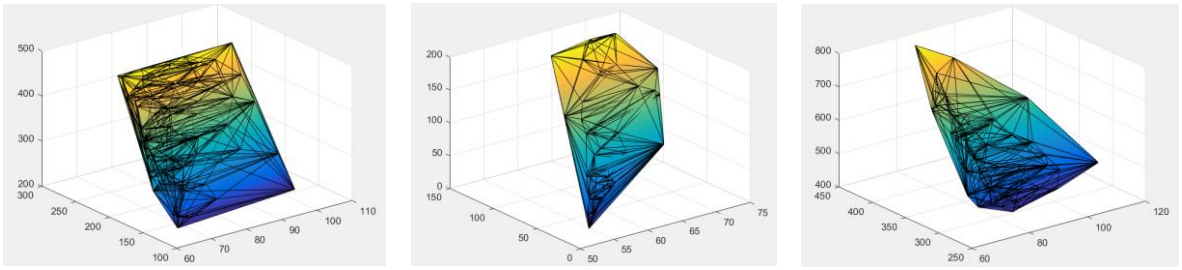
K-Means Clustering is required in the classification process of sound features. The sound feature in classification is based on the type of spoken word so that the clusters formed in the research there are 3. The cluster to be formed is a morning cluster representing the morning word, the cluster certi represents the word certification of appreciation and the techno cluster represents

the word information technology. Voice classification is used to test the stability of each speech's sound characteristics. Sound feature visualization after clustering is shown in Fig. 12. After classification, there are differences in the member cluster of each word spoken that cause a difference in voice recognition of the training data. The voice recognition result is shown in Table 3.

Table 3. Voice recognition of the training data

File Name	Morning	Certification of Appreciation	Information Technology
Morning1.m4a	✓		
Morning2.m4a	✓		
Morning3.m4a			✓
Certi5.m4a		✓	
Certi7.m4a		✓	
Certi8.m4a		✓	
Techno3.m4a			✓
Techno5.m4a			✓
Techno9.m4a			✓

Testing by using the sound feature of the test data. The sound test Data is obtained from 3 respondents who will pronounce the word in the morning, said certification of appreciation, and said information technology. There are 9 sample test data sounds that will be tested with members of the sound feature to all the clusters that have been formed. The result of the test is voice recognition by the system based on the number of members in a cluster shown in Table 4.



(a) morning cluster (b) certi cluster (c) techno cluster

Fig. 12. Visualization of clustering results

Table 4. Voice recognition of the testing data

File Name	Morning	Certification of Appreciation	Information Technology
Morningtest1.m4a	✓		
Morningtest2.m4a	✓		
Morningtest3.m4a	✓		
Certitest1.m4a		✓	
Certitest2.m4a		✓	
Certitest3.m4a		✓	
Technotest1.m4a			✓
Technotest2.m4a			✓
Technotest3.m4a	✓		

After getting the test results, the FAR, FRR, and Accuracy obtained refer to equation 11-13, so the results shown in Table 5 are obtained.

Table 5. Test results

Test	Result
FAR	0.94
FRR	0.11
Acc	92.59 %

IV. Conclusion

The results showed that the voice characteristics were built from the best training data selection results using correlation coefficients to get the best 3 sura data for each category. Voice feature extraction is done using imaginary numbers and real numbers formed from the fast Fourier transform stage. Sound features are visualized in 3-dimensional shapes that have white, gray, and black space patterns. Sound feature testing is performed using the K-Means method, which forms 3 clusters based on speech, namely the morning cluster, certi cluster, and techno cluster. The accuracy rate that was identified from 9 test data with different people's voices was 92.59%

Acknowledgment

The authors would like to express heartfelt thanks to The Modern Computing Research Center, Department of Information Technology, Politeknik Negeri Samarinda, for giving all their support.

References

[1] B. Dave and P. D. S. Pipalia, "Speech Recognition: a Review," *Int. J. Adv. Eng. Res. Dev.*, vol. 1, no. 12, pp. 230–236, 2014, doi: 10.21090/ijaerd.011244.

[2] K. R. Ghule and R. R. Deshmukh, "Feature-Extraction-Techniques-for-Speech-Recognition-A-Review.docx," *Int. J. Sci. Eng. Res.*, vol. 6, no. 5, pp. 143–147, 2015.

[3] M. Ference and A. M. Weinberg, "Center of Gravity and Center of Mass," *Am. J. Phys.*, vol. 6, no. 2, pp. 106–106, 1938, doi: 10.1119/1.1991277.

[4] Y. A. Ibrahim, J. C. Odiketa, and T. S. Ibiyemi, "Preprocessing technique in automatic speech recogntion for human computer interaction: an overview," *Ann. Comput. Sci. Ser.*, vol. XV, no. 1, pp. 186–191, 2017.

[5] A. G. Jondya and B. H. Iswanto, "Indonesian's Traditional Music Clustering Based on Audio Features," *Procedia Comput. Sci.*, vol. 116, pp. 174–181, 2017, doi: 10.1016/j.procs.2017.10.019.

[6] O. Of and E. For, "PCA- Based P Almprint R Ecognition 1 Introduction 2 The Structure of palmprint verification systems 3 Feature normalization techniques," *Electr. Eng.*, no. i, pp. 2–5, 2009.

[7] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.

[8] O. K. Hamid, "Frame Blocking and Windowing Speech Signal," *J. Information, Commun. Intell. Syst.*, vol. 4, no. 5, 2019.

[9] H. Triwiyanto, O Wahyunggoro, H A Nugroho, "Performance Analysis of the Windowing Technique on Elbow Joint Angle Estimation Using Electromyography," *J. Phys.*, 2018.

[10] H. Hauser, E. Gröller, and T. Theußl, "Mastering Windows: Improving Reconstruction," *2000 IEEE Symp. Vol. Vis. VV 2000*, pp. 101–109, 2000, doi: 10.1109/VV.2000.10002.

[11] R. Hibare and A. Vibhute, "Feature Extraction Techniques in Speech Processing: A Survey," *Int. J. Comput. Appl.*, vol. 107, no. 5, pp. 1–8, 2014, doi: 10.5120/18744-9997.

[12] A. K. . F. Haque, "FFT and Wavelet-Based Feature Extraction for Acoustic Audio Classification," *Int. J. Adv. Innov. Thoughts Ideas*, pp. 1–7, 2012.

[13] A.B.W. Putra, S. Pramono, and A. Naba, "Rancang Bangun Prototype Ciri Citra Kulit Luar Kayu Tanaman Karet Menggunakan Metode Virtual Center of Gravity," *J. EECCIS*, vol. 8, no. 1, p. pp.19-26, 2014.

[14] A. V. D. Sano and H. Nindito, "Application OF K-means algorithm for cluster analysis on poverty of provinces in indonesia," *ComTech*, no. 6, pp. 141–150, 2011.

[15] O. Oyelade, Oladipupo, "Application of k-Means Clustering algorithm for prediction of Students ' Academic Performance," *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, pp. 292–295, 2010.

[16] S. Saito, Y. Tomioka, and H. Kitazawa, "A Theoretical Framework for Estimating False Acceptance Rate of PRNU-Based Camera Identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 9, pp. 2026–2035, 2017, doi: 10.1109/TIFS.2017.2692683.