Enhanced blocking block area method for segmentation of continuous speech

B Arasyi, S N Endah, R Kusumaningrum, S Adhy
Department of Informatics, Diponegoro University
Prof. Soedharto SH., Tembalang Semarang

Corresponding author: sukmane@lecturer.undip.ac.id

Abstract. Segmentation is one of the important stages in the speech recognition in the type of continuous speech. The segmentation phase serves to break the sentences into words that can be recognized by the computer. The quality of segmentation results can affect the results of the recognition that is done. This research examines the dynamic threshold used in the process of continuous speech segmentation and also proposes an Enhanced Blocking Block Area method in the Indonesian language domain. Three algorithms were compared (K-Means, Fuzzy C-Means, and Otsu) to find the best dynamic threshold and add morphological operations and overlapping to the blocking block area method to obtain the best segmentation accuracy. Based on the results of the research, the Fuzzy C-Means algorithm provides the best threshold results compared to the other two algorithms. By using the Fuzzy C-Means algorithm with the addition of morphology and overlapping, this study can improve the accuracy of continuous speech segmentation in the Indonesian Language from 24% to 90%.

1. Introduction
Speech recognition is one example of the development of technological applications in sound media. Many factors influence the success of a speech recognition application, including health, age and gender of the examiners and the algorithms used in the application. The development of speech recognition (speech to text) has run quite rapidly; many studies have sprung up discussing speech recognition [1][2][3]. However, these four studies only discuss speech recognition that is applied to isolated words.

Speech recognition of isolated word types is speech recognition that can recognize individual words, not part of a single sentence, while speech recognition in the type of continuous speech is speech recognition that can recognize words as part of a sentence that is spoken naturally [4]. In continuous speech recognition applications, there are several stages, namely pre-processing, feature extraction, and recognition [5]. Pre-processing is a stage where the sound will be processed in such a way that it is ready for feature extraction. Pre-processing itself has several stages and among them is segmentation. The segmentation process functions as a sentence breaker into words.

Research related to segmentation has been done with several methods, including using dynamic thresholding and blocking block area methods that are applied in Hindi [6]. The proposed research will improve this method by adding two stages of the process that can improve its performance. The language domain for evaluation uses Indonesian.

In Rahman and Bhuiyan method, speech recognition in the type of continuous speech will be formed in a spectrogram image. Then, the spectrogram image will be changed in the form of binary images. To change it, of course, it takes a threshold that can dynamically be obtained through certain algorithms.
such as K-Means, Fuzzy C-Means and Otsu. The binary image results are then made into word blocks based on the number of pixels per column. Each word block is the result of its segmentation [6].

However, when this method was applied in the Indonesian language domain, the results were less than optimal. Thus, this study proposes an Enhanced Blocking Block Area Method by applying the morphological process in the binary image of the spectrogram and applying the concept of the overlapping column to determine a word block. This study will also compare the K-Means algorithm, Fuzzy C-Means and Otsu as dynamic thresholding in binary image spectrogram formation to get the threshold using classification. K-Means algorithm and Fuzzy C-Means are classification method that is often used and the result is optimal, while Otsu is a simple classification based on image class. The methods used to obtain various thresholds and the minimum result will be the threshold for segment detection.

2. Method

There are four important stages in the segmentation process, namely spectrogram, dynamic thresholding, blocking block area, and boundary detection and sound segmentation. The spectrogram converts the sound signal into an image of the intensity of the sound signal. The spectrogram image of the sound signal is converted into a binary form using the threshold resulting from the dynamic thresholding process. Binary image results from dynamic thresholding experience blocking block area processes, as well as boundary detection and sound segmentation. The purpose of the process is to convert the binary image into a beam shape so that it can determine the limit of each word. Based on each word's limit, the voice signal in the form of sentences is segmented into the form of each word. The description of the segmentation process flowchart can be seen in Figure 1. The following is a detailed explanation of each process.

2.1. Spectrogram

Examples of sound signal images, spectrogram images, and grayscale spectrogram images can be seen in Figure 2. The spectrogram box in the flowchart contains one process, while the dynamic thresholding process box contains process and sub-process.

Figure 1. Flowchart of segmentation process

The spectrogram functions to convert the input sound signal into a signal of a spectrogram image with a different signal density. Therefore, the spectrogram can be used to identify and classify input
sounds phonemically. The image of the spectrogram results is converted to a grayscale image which 
will then be processed into binary images in the dynamic thresholding process.

![Sound signal](image1)

![Spectrogram image](image2)

![Grayscale spectrogram images](image3)

**Figure 2.** Sound signal, Spectrogram images, dan Grayscale spectrogram images signal “it has two 
black cars”

2.2. Dynamic thresholding
To convert a grayscale image into a binary image, a threshold value is needed. This study 
compares the results of three algorithms, namely K-Mean, FCM, and Otsu to find the best threshold value so that it 
can minimize errors during word cutting.

2.2.1. K-Mean. The K-Mean algorithm is one of the most well-known clustering algorithms. The purpose 
of this algorithm is to divide the data into several groups. Data grouping is first done randomly. The 
centroid is re-estimated until no new centroid is found. In this study, researchers used the K-Mean 
algorithm with three centroids and calculated the average of the first and second-largest centroids. This 
average result will be used as a dynamic thresholding. The formulation for calculating the threshold of 
the sound spectrogram is as follows [6]:

\[
J = \sum_{i=1}^{k} \sum_{j=1}^{c_i} (\|z_i^{(j)} - v_j\|^2)
\]

where:
\[
\|z_i^{(j)} - v_j\| = \text{euclidean distance between data (}z_i\text{) and cluster center (}v_j\text{)}.
\]
\[
c_i = \text{the number of data in cluster i}.
\]
\[
k = \text{the number of the cluster center}.
\]
K-mean algorithm has six algorithm steps, including [8]:

1. Determine the number of clusters, for example, the number of clusters is three, \( k = 3 \). Thus, \( \mathbf{v} = \{v_1, v_2, v_3\} \) are three sets of cluster centers.
2. Calculate the distance \( d_{ij} \) between data \( i \) and cluster center \( j \).
3. Cluster each data based on the closest cluster center according to the minimum distance.
4. Re-calculating the new cluster center with the formula as follows:
   \[
   v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_j
   \]
   where:
   - \( v_i \) = cluster centers.
   - \( x_j \) = data index \( j \).
   - \( c_i \) = number of data in cluster \( i \).
5. Re-calculating the distance between each data and the newly acquired cluster center.
6. If there is no new data to be clustered, then stop. But, if there is, go back to the third step.

2.2.2. Fuzzy C-Mean. Fuzzy C-Mean (FCM) is a general form of the K-Mean algorithm. If in the K-Mean Algorithm, the data will be grouped based on one nearest centroid, then the FCM data will be grouped into all centroids with certain membership degrees. Just like K-Mean, in this study researchers used three centroids, which would later be searched for on average from the two largest centroids. This average result will be used as a dynamic thresholding. There are six steps in the FCM algorithm, which are as follows [8]:

1. Determine the number, centroid(\( c \)) = 3 and the cluster center value. Suppose that is a cluster center set.
2. Calculate the Euclidean distance \( d_{ij} \) between data \( i \) and cluster center \( j \).
3. Re-newing the fuzzy membership system \( (\mu_{ij}) \) with the following formula:
   \[
   \mu_{ij} = \frac{1}{\sum_{k=1}^{c}(d_{ij}/d_{ik})^{m-1}}, \forall i = 1,2,...,n; \forall j = 1,2,...,c
   \]
   where:
   - \( \mu_{ij} \) = new fuzzy
   - \( d_{ij} \) = euclidean distance between data \( i \) and cluster center \( j \)
   - \( d_{ik} \) = euclidean distance between data \( i \) and cluster centers
   - \( m \) = fuzzy index \([1,\infty]\)
4. Calculate the fuzzy center ‘\( v_j \)’ by using:
   \[
   v_j = \left( \frac{\sum_{i=1}^{n}(\mu_{ij})^m x_i}{\sum_{i=1}^{n}(\mu_{ij})^m} \right), \forall j = 1,2,3,...,c
   \]
   where:
   - \( v_i \) = cluster centers.
   - \( \mu_{ij} \) = new fuzzy
   - \( x_j \) = data index \( j \).
   - \( m \) = the fuzzy index \([1,\infty]\).
5. Iterate from the second step until the minimum value ‘\( J \)’ is acquired or \( \| U^{k+1} - U^k \| < e \).
   where:
   - \( k \) = iteration step.
   - \( e \) = termination criterion between \([0,1]\).
   - \( U = (U_{ij})_{n \times c} \) = fuzzy membership matrix, dan \( J \) is an objective function, obtained with the following formula:
   \[
   J = \sum_{j=1}^{c} \sum_{i=1}^{n}(\mu_{ij})^m d_{ij}
   \]
6. Calculate the dynamic thresholding by calculating the average of the two largest centroids.
2.2.3. Otsu’s method. The purpose of the Otsu's method is to divide the grayscale image histogram into two different areas automatically without requiring user help to enter a threshold value. The approach taken by Otsu’s method is to do discriminant analysis which is to determine a variable that can distinguish between two or more groups that appear naturally. The discriminant analysis will maximize these variables in order to divide the foreground and background objects [9].

Otsu’s algorithm assumes that images are composed of two basic classes: Foreground and Background. The calculation is done to find the optimal threshold value that minimizes the variance value in the class (within-class variance) but also maximizes the variance value between classes (between-class variance) [6].

The mathematical formula of the Otsu’s method for calculating the optimum threshold is as follows [10]:

For example, \( P(i) \) is histogram image from a voice. Probability of two classes \( w_1(t) \) and \( w_2(t) \) on level \( t \) can be counted as follows:

\[
\text{\( w_1(t) = \sum_{i=1}^{t} P(i) \) and \( w_2(t) = \sum_{i=t+1}^{t} P(i) \)}
\]

where:

- \( P(i) \) = histogram image

Class average, \( \mu_1(t) \) and \( \mu_2(t) \) are:

\[
\mu_1(t) = \sum_{i=1}^{t} \frac{ip(i)}{w_1(t)} \text{ and } \mu_2(t) = \sum_{i=t+1}^{t} \frac{ip(i)}{w_2(t)}
\]

where:

- \( \mu(t) \) = average of the class

Class variant:

\[
\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{w_2(t)}
\]

\[
\sigma_2^2(t) = \sum_{i=t+1}^{t} [i - \mu_2(t)]^2 \frac{P(i)}{w_2(t)}
\]

where:

- \( \sigma^2(t) \) = variant of class

The variant value in class \( \sigma_w \) is counted by adding the value of variant from those two classes using the following formula:

\[
\sigma_w^2 = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t)
\]

where:

- \( \sigma_w^2 \) = within-class variant value

\[
\text{\( \sigma^2(t) \) = variant of class}
\]

\[
\text{\( w(t) \) = weight of probability}
\]

The variant value between class \( \sigma_b \) is determined by calculating the margin from the variant total and the variant value of the class, which can be acquired using the following formula:

\[
\sigma_b^2 = \sigma^2(t) - \sigma_w^2(t) = w_1(t)w_2(t)[\mu_1(t) - \mu_2(t)]^2
\]

where:

- \( \sigma_b^2 \) = between-class variant value

- \( \sigma^2(t) \) = variant of class

- \( w(t) \) = weight of probability

- \( \mu(t) \) = average of the class

These two variants \( \sigma_w \) and \( \sigma_b \) are counted for all possibilities of the threshold, \( t=0…1 \) (maximum intensity).
2.3. **Morphology operation**

In this study, two morphological operations are used including erosion and dilatation. The concept of erosion and dilation operations can be explained as follows [11].

2.3.1. **Erosion.** In the erosion process, binary images are divided into two, namely foreground and background. This process will match whether there are parts of the element structure that come into contact with the background when the center of the element structure is in contact with the foreground; if there is, then the foreground value should be changed according to the background value.

2.3.2. **Dilation.** The dilation process applies the same concept as the erosion process; it's just that this process is the opposite of the erosion process. This process will match whether there are parts of the element structure that come into contact with the background when the center of the element structure is in contact with the foreground; if any, the background value that comes into contact with the element structure should be changed according to the foreground value.

2.4. **Forming of block area**

The blocking block area algorithm functions to convert spectrogram binary images into block images. Researchers used the blocking block area method that has been modified by applying the concept of overlapping columns. The course of the blocking block area algorithm by dividing the image into several frames and calculating the number of black values (luminance 0) and the white value (luminance 1) in each frame. The frame will be marked in black if the black value reaches 45% or more than the number of pixels in the frame and white if the white color reaches 55% or more. Change all pixel values in the frame to white if the frame mark and the frame after that are white, otherwise change the entire pixel value in the frame to black.

The steps in the 'Blocking Block Area' method are as follows [6]:
1. Add the intensity value of the binary image in each column
2. Change the value of the column to black (luminance 0), if the white pixel value is less than the black pixel value.
3. Change the column value to white (luminance 1), if the white pixel value is more than the black pixel value.
4. If the sign of a frame is white and the next frame is white then change the frame to white (luminance 1). In addition, change the frame value to black (luminance 0).
5. Repeat Step 5 until all frames change color.

However, the method applied by Rahman & Bhuiyan is not appropriate if applied to the Indonesian language domain, therefore in this study, researchers tried to improve it by applying the concept of the overlapping column in the blocking block area method to obtain maximum results. Following is the application of the overlapping column concept in the blocking block area method.

1. Divide the morphological image into several frames with a length and a certain overlap value.
2. Add the value of each pixel in each frame.
3. Mark the frame with a black value, if the white pixel value is less than the black pixel value (luminance 0).
4. Mark the frame with a white value, if the white pixel value is more than the black pixel value (luminance 1).
5. Repeat Step 5 until all frames change color.

Figure 3, Figure 4, and Figure 5 respectively are binary images from dynamic thresholding results, binary images that have gone through morphological processes, and blocking block area results to morphological results.
2.5. Boundary detection and voice segmentation

The block image that is available is further processed at this stage. The coordinates of the initial limit and the end of each block are searched and the percentage value is calculated against the overall value of the block image column.
Figure 6. The example of start and end limits of block image

The percentage value of the initial limit and the deadline are used as guidelines for cutting votes.

Figure 7. The example of sound signal cutting

3. Experiment results

3.1. Segmentation testing scenario
In segmentation testing, testing was carried out on two different methods, namely the method in [6] and the method that was developed, namely with the addition of morphological processes, and conceptualization of the overlapping column to determine the type of block in the blocking block area. The purpose of this test is to determine the accuracy of each method in the Indonesian domain and to find out the best algorithm (K-Means, FCM, Otsu) used at dynamic thresholding.
In segmentation testing, 80 data were used consisting of 20 different sentences spoken by four different people Person1 (P1), Person2 (P2), Person3 (P3), and Person4 (P4) of the male sex. The sample selection is based on the pronunciation accent in Java Island.

3.2. Segmentation testing results
Tables 1 and 2 are the tables of the segmentation testing results using the Rahman & Bhuiyan method, and the method proposed by the researcher. To get accurate results, the following formula is used.

\[
\text{Accuracy} = \frac{\text{number of words segmented correctly}}{\text{number of words}} \times 100\% \tag{12}
\]
The numbers 0 – 5 in the results are obtained by calculating the number of words segmented correctly in each sentence.

**Table 1. Segmentation results using rahman & bhuiyan method**

| No | Sentence | Number of words | K-means | Number of words segmented correctly | Otsu |
|----|----------|----------------|---------|-------------------------------------|------|
|    |          |                | P1      | P2      | P3      | P4      | P1      | P2      | P3      | P4      | P1 | P2 | P3 | P4 |
| 1  | K1       | 5              | 1       | 2       | 0       | 0       | 1       | 2       | 0       | 0       | 0  | 0  | 0  | 0  |
| 2  | K2       | 5              | 4       | 1       | 2       | 1       | 4       | 1       | 2       | 1       | 0  | 0  | 1  | 0  |
| 3  | K3       | 5              | 1       | 2       | 0       | 0       | 1       | 1       | 0       | 0       | 0  | 0  | 0  | 0  |
| 4  | K4       | 5              | 3       | 3       | 3       | 0       | 3       | 3       | 1       | 0       | 0  | 0  | 0  | 0  |
| 5  | K5       | 5              | 0       | 2       | 0       | 1       | 0       | 2       | 0       | 0       | 0  | 0  | 0  | 0  |
| 6  | K6       | 5              | 4       | 2       | 1       | 0       | 2       | 2       | 1       | 0       | 0  | 0  | 0  | 0  |
| 7  | K7       | 5              | 2       | 3       | 1       | 1       | 2       | 2       | 1       | 1       | 0  | 0  | 0  | 0  |
| 8  | K8       | 5              | 2       | 0       | 0       | 0       | 2       | 0       | 0       | 0       | 0  | 0  | 0  | 0  |
| 9  | K9       | 5              | 3       | 3       | 1       | 1       | 3       | 1       | 1       | 0       | 0  | 0  | 0  | 0  |
| 10 | K10      | 5               | 0      | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0  | 0  | 0  | 0  |
| 11 | K11      | 5               | 2       | 1       | 2       | 2       | 2       | 1       | 1       | 2       | 0  | 0  | 0  | 0  |
| 12 | K12      | 5               | 3       | 0       | 2       | 1       | 1       | 0       | 1       | 1       | 0  | 0  | 0  | 0  |
| 13 | K13      | 5               | 1       | 1       | 1       | 2       | 1       | 1       | 1       | 2       | 0  | 0  | 0  | 0  |
| 14 | K14      | 5               | 3       | 3       | 3       | 3       | 3       | 3       | 3       | 3       | 2  | 0  | 1  | 0  |
| 15 | K15      | 5               | 3       | 0       | 0       | 2       | 1       | 1       | 0       | 1       | 0  | 0  | 0  | 1  |
| 16 | K16      | 5               | 3       | 3       | 1       | 2       | 2       | 1       | 1       | 1       | 0  | 0  | 0  | 0  |
| 17 | K17      | 5               | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0  | 0  | 0  | 0  |
| 18 | K18      | 5               | 4       | 3       | 1       | 2       | 2       | 2       | 1       | 1       | 0  | 0  | 0  | 0  |
| 19 | K19      | 5               | 2       | 4       | 3       | 3       | 2       | 3       | 3       | 3       | 0  | 1  | 0  | 0  |
| 20 | K20      | 5               | 2       | 2       | 1       | 2       | 2       | 1       | 1       | 2       | 1  | 1  | 0  | 1  |

**Average of Accuracy** 30% 24% 2%

Where:

$P_i$, $i = 1 \ldots 4$ shows the people.

$K_j$, $j = 1 \ldots 20$: shows the sentences.

**Table 2. Results of proposed segmentation method**

| No | Sentence | Number of words | K-means | Number of words segmented correctly | Otsu |
|----|----------|----------------|---------|-------------------------------------|------|
|    |          |                | P1      | P2      | P3      | P4      | P1      | P2      | P3      | P4      | P1 | P2 | P3 | P4 |
| 1  | K1       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 4       | 5       | 4  | 4  | 1  | 4  |
| 2  | K2       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 0  | 2  | 3  | 4  |
| 3  | K3       | 5              | 5       | 5       | 5       | 2       | 4       | 5       | 5       | 2       | 4  | 3  | 4  | 3  |
| 4  | K4       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 3  | 3  | 5  | 3  |
| 5  | K5       | 5              | 5       | 2       | 5       | 5       | 3       | 5       | 5       | 2       | 4  | 1  | 4  | 4  |
| 6  | K6       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 2  | 4  | 1  | 4  |
| 7  | K7       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 2  | 4  | 3  | 3  |
| 8  | K8       | 5              | 5       | 3       | 3       | 3       | 5       | 5       | 5       | 3       | 2  | 3  | 2  | 0  |
| 9  | K9       | 5              | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 2  | 3  | 3  | 4  |
| 10 | K10      | 5              | 3       | 1       | 3       | 5       | 5       | 5       | 5       | 5       | 2  | 2  | 1  | 3  |
| 11 | K11      | 5              | 5       | 2       | 3       | 4       | 5       | 2       | 3       | 4       | 3  | 4  | 2  | 1  |
| 12 | K12      | 5              | 5       | 3       | 5       | 5       | 3       | 3       | 5       | 2       | 4  | 5  | 5  | 5  |
Based on Tables 1 and 2, the percentage of recapitulation from segmentation results can be seen in Table 3 below. BBAM represents the Blocking Block Are Method and BBAM+MO+OV represent the Blocking Block Are Method + Morphology Operation + Overlapping which has been explained in the methodology section.

Table 3. Segmentation results percentage

| No | Thresholding Algorithm | Average of Accuracy | Detail |
|----|-------------------------|--------------------|--------|
|    |                         | BBAM               | BBAM+MO+OV |        |
| 1  | K-Mean                  | 30%                | 88%       | Increase 58% |
| 2  | FCM                     | 24%                | 90%       | Increase 66% |
| 3  | Otsu                    | 2%                 | 60%       | Increase 58% |

3.3. Segmentation results analysis

It can be seen in Table 3 above that dynamic thresholding using the FCM algorithm provides the best results when compared to using the K-Means and Otsu algorithms. Segmentation results using the BBAM method are applied to the Indonesian domain the results are less accurate. This could be due to phonemic differences between Hindi and Indonesian. Therefore, improvements were made to the BBAM method (BBAM + MO + OV) so as to produce the highest accuracy by 90% and can increase segmentation accuracy by 66%.

The application of the morphological process to segmentation causes the sound segmentation in the blocking block area to be uneven because there is a voice signal that is not part of the word, becomes segmented.

As seen in Figure 8 (a), the red circle in the picture above shows that the voice signal that is not part of the word becomes segmented and causes voice on the voice of segmentation.
4. Conclusion
The best algorithm for finding the threshold value in this study is the Fuzzy C-Mean algorithm (FCM). The method proposed in this study adds morphological processes to binary spectrogram images and applies the concept of overlapping columns to determine the type of block in the Blocking Block Area stage so as to be able to increase accuracy in the Indonesian language domain by 66%. The level of accuracy of segmentation by using the method proposed by the researcher reached 90%.

Acknowledgment
The authors would like to thank KemenristekDikti for supporting this research.

References
[1] S N Endah, S Adhy and Sutikno 2017 Telkomnika 15 292–8
[2] S N Endah, P S Sasonko, H A Wibawa, Sutikno, and R Kusumaningrum 2017 Adv. Sci. Lett. 23 6555–8
[3] H Park 2009 Integrated Phoneme Subspace Method for Speech Feature Extraction Eurasip 2009 1-6
[4] S K Gaikwad, B W Gawaliand P Yannawar 2010 International Journal of Computer Applications 10 16–24
[5] S N Endah, S Adhy and Sutikno 2015 Telkomnika 13 571–7
[6] M M Rahman and M A A Bhuiyan 2013 International Journal of Research in Engineering and Technology 2 404–11
[7] N Dhanachandra, K Manglem and Y J Chanu 2015 Procedia Computer Science 54 764–71
[8] J Nayak, B Naik and H Behera 2015 Proceedings of the International Conference on CIDM (Computational Intelligence in Data Mining vol 2) (India: Odisha) pp
[9] Review on Otsu Image Segmentation Algorithm International Journal of Advanced Research in Computer Engineering & Technology 2 387–9
[10] M M Rahman, F Khatun and M AA Bhuiyan 2015 International Journal of Advanced Research in Artificial Intelligence 4 1–6
[11] J Naser and S Nanik 2013 International Journal of Advanced Science and Technology 51 128–34