Automatic speech recognition by using local adaptive thresholding in continuous speech segmentation

S N Endah, R Kusumaningrum, S Adhy, R A Ulfattah
Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia
Corresponding author: sukmane@lecturer.undip.ac.id

Abstract. Speech recognition of continuous speech will be greatly influenced by the word segmentation process of the speech input. Proper segmentation will result in better speech recognition. This study proposed automatic speech recognition by applying local adaptive thresholding in the segmentation process. The segmentation method used is an enhanced blocking block area method whose input is a spectrogram image of the speech signal. While the locally adaptive thresholding method used is the Niblack method which is the best method compared to other methods, namely Sauvola, Bradley, Guanglei Xiong, and Bernsen when applied to the enhanced blocking block area method. For the speech recognition process, using Mel-frequency cepstral coefficients (MFCC) as a feature extraction method and Hidden Markov Model (HMM) as speech recognition. The experimental results show that by using 400 sentences consisting of 80 testing data and 320 as training data and using K-fold cross-validation, the highest accuracy is 60.8%. This result has no significant difference with the use of global thresholding in the segmentation process.

1. Introduction
Continuous speech is a series of words continuously without a clear boundary between words, like human speech. The recognition of continuous speech types will be very useful, especially in today's digital era. Several researchers have made various applications for the development of speech recognition [1]. The rapid development of technology has triggered the development of various intelligent systems that accept speech input. So that proper speech recognition is the target of researchers to always experiment with various existing methods.

Continuous speech recognition is the development of isolated word recognition, which recognizes words from a sentence using machine learning algorithms [2]. The stages of developing continuous speech recognition include pre-processing, feature extraction, and recognition.

The pre-processing stage is the stage for preparing speech signals so that feature extraction can be carried out. One of the stages in pre-processing is the segmentation process. Segmentation is dividing continuous speech into basic units such as words, phonemes, or recognizable syllables [3]. Continuous speech segmentation can be done by converting the speech signal representation into a spectrogram image that is processed into word segments, such as the processes carried out in the Blocking Block Area method [4] and the Enhanced Blocking Block Area [5, 6].

The blocking Block Area method is the process of making word blocks from a spectrogram image into a binary image by going through several stages, namely changing the speech signal representation into a spectrogram, performing dynamic thresholding with a clustering algorithm on the spectrogram
image to produce a binary image then carrying out boundary detection using the Blocking method.

Block Area. Then this research was improved by adding morphological operations before the Blocking Area Blocking was carried out and overlapping the block-making process, which was then called Enhanced Blocking Block Area [5]. The process of forming a speech signal spectrogram image into a binary spectrogram image using the Blocking Block Area method in the research conducted by [4] and [5] uses dynamic thresholding to determine the threshold of the spectrogram image globally.

In determining the global dynamic threshold, a single threshold is generated based on the entire pixel information of the spectrogram image [7]. The threshold can work well on images containing objects with uniform intensity values on a contrasting background, but errors will occur if there is low contrast (illumination) between the object and the background or noise in the image [7]). The background in a spectrogram image has varying illumination and contrast so that an error will occur when using the global threshold to differentiate between object and background. Therefore, good thresholding is needed for degraded or illuminated images, namely using Local Adaptive Thresholding techniques [7].

Local Adaptive Thresholding is a technique for generating binary images by estimating different threshold values for each pixel according to grayscale information with local statistics from neighboring pixels [8]. Local Adaptive Thresholding uses the window to limit the region to be determined by the threshold for each image pixel in the window with a certain size. The use of this window will better represent the overall pixel value of the image so that it can obtain a good threshold value on degraded images or non-uniform images [9]. There are several methods in Local Adaptive Thresholding [10], including Niblack, Sauvola, Bradley, Guanglei Xiong Statistical Thresholding and Bernsen. Niblack, Sauvola, and Bernsen have an accuracy value of> 75% in binarization of documents with illumination [11], while Bradley, Guanglei Xiong Statistical Thresholding has the fastest average processing time of the five methods in carrying out binarization, but the accuracy value is not discussed [10].

Research conducted by Ulfattah et al. (2020) tries to improve the Enhanced Blocking Block Area method by using Local Adaptive Thresholding in forming a speech signal spectrogram image into a binary spectrogram image [6]. Several methods in Local adaptive thresholding are compared to determine the appropriate method in carrying out word segmentation in the Indonesian domain. These methods are Niblack, Sauvola, Bradley, Guanglei Xiong Statistical Thresholding, and Bernsen. The results of this study indicate that the level of accuracy to be able to segment words correctly is 95% using the thresholding method of Niblack [6].

This study discusses automatic speech recognition by applying local adaptive thresholding, namely the Niblack method in segmenting words using Enhanced Blocking Block Area. For the feature extraction stage using the MFCC method and the introduction stage using HMM.

2. Methods
The research methodology has 6 main stages, namely Data Collection, Data Sharing, Pre-Processing, Feature Extraction, Introduction, and Evaluation. For more details, it can be seen in Figure 1.
The details of each process in the methodology can be explained as follows.

2.1. Data collection
The process to get voice data is done by recording sound using the Matlab program. Five different people were recorded for 20 sentences of continuous speech, each sentence consisting of five words at a frequency of 44100 Hz. Each person says the same sentence 4 times with different intonations. The results of the recording are then saved in *.wav format. The total data collected was 400 sentences.

2.2. Data partition
The data of the continuous speech voice signal from the recording is then divided to separate the data for training and testing. This is so that more accurate model test results are obtained. Data sharing is done using 5-fold cross-validation so that the percentage is 20% for testing data and 80% for training data.
2.3. Pre-processing
In this pre-processing, there are 2 stages, namely segmentation, and normalization.

2.3.1. Segmentation. The segmentation process aims to break down continuous speech, which is initially in the form of a sentence, into words. There are five processes in continuous speech segmentation, namely, Generate Spectrograms, Binarization Using Local Adaptive Thresholding, Morphological Operations, Improved Blocking Block Areas, and Boundary Detection [6]. The local adaptive thresholding used is the Niblack method. Niblack determines the threshold value based on the local mean and local standard deviation. Both are calculated in a window with the size of m x n based on the neighborhood value of the pixels so that each pixel has a different threshold value. The formula to calculate the threshold value is [6, 8].

\[ T(i,j) = m(i,j) + k \cdot \sigma(i,j) \]  

where:
- \( k \): a constant that has a value between 0 and 1
- \( m(i,j) \): the local mean of the pixel in the local window
- \( \sigma(i,j) \): the local standard deviation of the pixel in the local window

2.3.2. Normalization. The sound signal generated from the segmentation process is then carried out with the normalization process with the DC-removal algorithm. Normalization is done by calculating the average of the sample signal and subtracting the average value of each sample signal.

2.4. Feature extraction
The method used for this is the Mel Frequency Cepstral Coefficients (MFCC). The stages of the process for the MFCC are as follows: Pre-emphasize, Frame Blocking, Windowing, FFT, Mel-Frequency Wrapping, Mel-Frequency Cepstrum, and cepstral Filtering. Preemphasize serves to reduce the signal-noise ratio and balance the spectrum of the sound of a voice. The Frame Blocking Process is used to cut the sound signal of long duration becomes shorter duration in order to get characteristic of the periodic signal. The windowing process aims to reduce spectral leakage or aliasing, which is the effect of the blocking frame which causes the signal becomes discontinue. FFT (Fast Fourier Transform) is a transformation method to get a signal in the frequency domain of the discrete signals that exist. Filterbank was conducted in order to determine the energy in the sound signal. The frequency of a signal is measured using a mel scale. Mel-Frequency Cepstrum obtained from DCT process to get back the signal in the time domain. The result is called the Mel-Frequency cepstral coefficient (MFCC). The results of MFCC have several drawbacks, namely the low-order, which is very sensitive to the spectral slope, and the high-order, which is very sensitive to noise. Therefore, the cepstral filtering has into one of these methods to minimize sensitivity [1, 12].

2.5. Recognition
This stage has 2 processes, namely Training and Testing. The method used is the Hidden Markov Model (HMM). The approach in the Hidden Markov Model (HMM) is to classify the spectral characteristics in each part of the sound in several patterns. The concept of HMM is to classify the sound signal as a random parametric process, and the parameters of this process can be recognized (predicted) with precise accuracy [12, 13]. The components of HMM are [1]:
1. Amount of state (N)
   The state is a hidden parameter or better known as a hidden state. In application amount of this state become one of thus testing parameter. So, the amount of state is set in such a way to obtain an optimal output. The number of states in the model Nstate labeled with \( S = \{ S_1, S_2, ..., S_N \} \).
2. Model Parameter (M)
Number of observation symbols that different in each state M. Observation symbol correlates with physical output from modeled system. Individual symbols are denoted by \( V = \{ v_1, v_2, v_3, \ldots, v_M \} \).

3. Early state distribution \( \pi = (\pi_i) \) where
\[
\pi = P[q_1 = i], 1 \leq i \leq N
\]  

4. Transition probability distribution state \( A = (a_{ij}) \) where
\[
a_{ij} = P[q_{u+1} = s_j | q_u = s_i], 1 \leq i, j \leq N
\]  That is probably an observation is in a state \( j \) when \( u+1 \) and when state \( i \) when \( u \).

5. The observation symbol probability distribution \( B = \{ b_j(k) \} \) where
\[
b_j(k) = P[o_u = v_k | q_u = j], 1 \leq k \leq M
\]  Represent symbol distribution in state \( j, j = 1, 2, 3, \ldots, N \)

According to five component above, to plan HMM, needs two model parameters that is \( N \) and \( M \), besides it also needs three possibilities \( (\pi, A, B) \) that is modeled by use notation \( \lambda \) \( [\lambda = (A, B, \pi)] \).

According to Rabiner [14], the problem can be solved by HMM are:
1. Arrange parameter \( \lambda = P(A, B, \pi) \) in order to produce maximum \( P(O | \lambda) \)
2. Counting \( P(O | \lambda) \) if known an observation sequence \( O = O_1, O_2, \ldots, O_T \) and a model \( \lambda = P(A, B, \pi) \).

2.6. Evaluation
Evaluation is done by calculating the accuracy of the results of the introduction.

3. Results and analysis

3.1. Data
The data used amounted to 400 data in the form of voice recordings of 20 sentences in Indonesian. The sentences can be seen in Table 1.

| No. | Sentence                                 | Sentence Code |
|-----|------------------------------------------|---------------|
| 1   | abang bercerita sesuatu yang bagus       | K1            |
| 2   | bapak ibu pergi bersama adik             | K2            |
| 3   | bibi mulai terkenal sore ini             | K3            |
| 4   | cincin kawin dari bahan permata          | K4            |
| 5   | dia punya dua mobil hitam                | K5            |
| 6   | hidup itu seperti sekotak coklat        | K6            |
| 7   | kamu jangan jadi judes juga              | K7            |
| 8   | kapan kita main bola pantai              | K8            |
| 9   | karena keju adalah susu sapi             | K9            |
| 10  | kompor kredit berwarna merah muda        | K10           |
| 11  | maaf atas kejadian senin lalu            | K11           |
| 12  | makan kuning telur setengah matang       | K12           |
| 13  | masinis kereta berbaju biru tua          | K13           |
| 14  | nanti siang saja kata berbahaya          | K14           |
| 15  | pabrik gula pasir ada lima               | K15           |
| 16  | paman meninggal saat dulu sekali         | K16           |
| 17  | pantun tentang pisang dan sayur          | K17           |
The twenty sentences were pronounced by five male genders (P1, P2, P3, P4, P5), each person uttered the same sentence 4 times.

3.2. Test scenario
The test uses k-fold cross-validation with k = 5. So that from 400 data is divided into 80 as test data and 320 as training data. The distribution of data for each fold is carried out based on the sentences spoken by each person. In the first fold, the test data are P1 speech data, and the training data are P2, P3, P4, and P5 speech data. Analogous to the first fold, the second to the fifth fold also applies. In testing this study using 3 scenarios with the following explanations.

Scenario 1. Perform continuous speech recognition testing with the segmentation process using local adaptive thresholding, as proposed in this study.
Scenario 2. Conducting continuous speech recognition testing with the segmentation process using global thresholding, as proposed by Arasyi (2020) [5], as the baseline in this study.
Scenario 3. Comparing the test results from scenario 1 and scenario 2.

The parameter used in scenario 1 and scenario 2 is a combination of the number of HMM states and the number of MFCC coefficients. The HMM states used are 14 and 15, and the MFCC coefficient used are 12 and 20.

3.3. Test result

3.3.1. Scenario 1. The results of scenario 1 can be seen in Table 2.

| Parameter | HMM State | MFCC Coeff. | Fold | Average Accuracy |
|-----------|-----------|-------------|------|------------------|
|           |           |             | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |            |
| 1         | 14        | 12          | 59%  | 69%  | 63%  | 58%  | 55%  | 60.8%      |
| 2         | 14        | 20          | 46%  | 66%  | 53%  | 52%  | 47%  | 52.8%      |
| 3         | 15        | 12          | 57%  | 71%  | 62%  | 57%  | 55%  | 60.4%      |
| 4         | 15        | 20          | 45%  | 69%  | 54%  | 51%  | 44%  | 52.6%      |

3.3.2. Scenario 2. The results of scenario 2 can be seen in Table 3.

| Parameter | HMM State | MFCC Coeff. | Fold | Average Accuracy |
|-----------|-----------|-------------|------|------------------|
|           |           |             | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |            |
| 1         | 14        | 12          | 53.7% | 67.25% | 64%  | 53.5% | 57.75% | 59.24%    |
| 2         | 14        | 20          | 47.5% | 67%  | 55.5% | 40.25% | 54%  | 52.85%    |
| 3         | 15        | 12          | 55.5% | 67%  | 61.5% | 52.25% | 65.25% | 60.3%     |
| 4         | 15        | 20          | 48.75% | 65.5% | 58.5% | 40.25% | 62%  | 55%       |
3.3.3. **Scenario 3.** In scenario 3 will be compared between scenario 1 and scenario 2 by looking at the Average Accuracy of each parameter. Figure 2 shows the results of scenario 3.

![Figure 2](image)

**Figure 2.** Comparison of scenario 1 and scenario 2.

### 3.4. Result analysis

Based on Figure 2, it can be seen that the highest accuracy is obtained when speech recognition is carried out using segmentation with local adaptive thresholding, which is 60.8%. This is achieved in the parameter HMM state 14 and MFCC coefficient 12. While speech recognition using segmentation with global thresholding has the highest accuracy of 60.3% in HMM state 15 and MFCC coefficient 12. So that the use of segmentation as proposed is only able to increase its accuracy by 0.5%. In other words, there is no significant difference for speech recognition either when segmenting with local adaptive thresholding or global thresholding, although the accuracy of the segmentation results is above 90%.

The low speech recognition accuracy can be caused by several factors, including differences in pronunciation, the surrounding conditions during the recording process, the sound quality, which is influenced by the health condition of the speakers, and the noise contained in the training data or test data. It can change the sound value.

Other factors that affect the results of speech recognition include imperfect word segmentation results, the quality of the microphone used, and the amount of training data used. In addition, it could be the choice of the MFCC method as feature extraction or the HMM method, which is not appropriate for the Indonesian continuous speech domain.

### 4. Conclusion

In this study, it can be concluded that the results of recognition using local adaptive thresholding is not significantly different from recognition using global thresholding and are only able to achieve the highest accuracy of 60.8%. This can be caused by several factors such as differences in pronunciation, surrounding conditions during the recording process, sound quality, which is influenced by the health condition of the speaker, the presence of noise or due to the selection of feature extraction and recognition methods which is incorrect. So, in future research, we can improve the quality of our speech database and try various other speech recognition methods such as deep learning which is currently trending.
Acknowledgment
The authors would like to acknowledge the research funding supported by Kementrian Pendidikan dan
Kebudayaan under the Grant of Fundamental Research for College Flagship – Contract Number 101-
50/UN7.6.1/PP/2020

References
[1] Endah S N, Adhy S and Sutikno 2017 *Telkomnika* 15 292-298
[2] Endah S N, Adhy S and Sutikno 2015 *Telkomnika* 13 571-577
[3] Rahman M and Bhuiyan A 2012 *Int. J. Adv. Comput. Sci. Appl.* 3 131-138
[4] Rahman M and Bhuiyan A 2013 *Int. J. Research in Engineering and Tech.* 2 404-411
[5] Rogowska J 2006 Digital Image Processing Techniques For Speckle Reduction, Enhancement, And Segmentation Of Optical Coherence To Monography (OCT) Images *Optical Coherence Tomography: Principles And Applications* (USA, Elsevier Inc) p 305-329a
[6] Al-Nasser M, Elshafei M and Al-Sarkhi A 2014 *Proc. of the ASME 2014 4th Joint US-
European Fluids Engineering Division Summer Meeting* (Chicago) 1 1-7
[7] Sauvola J and Pietika M 2000 *The J. of The Pattern Recognition Society* 33 225-236
[8] Arasyi B, Endah S N, Kusumaningrum R and Adhi S 2020 *J. Phys.: Conf. Ser.* 1524 012102
[9] Ulfattah R A, Endah S N, Kusumaningrum R and Adhy S, 2020 *Telkomnika* 18 407-418
[10] Londhe N D and Kshirsagar G B 2017 *Int. Conf. on Communication and Signal Processing,* (Tamilnadu, India)
[11] Chaki N, Shaikh S H and Saeed K 2014 Comprehensive Survey on Image Binarization Techniques *Exploring Image Binarization Techniques* (New Delhi, Springer, India) p. 5-15
[12] Singh T R, Roy S, Singh O I, Sinam T and Singh KM 2011 *Int. J. of Comp. Science Issues* 8 271–277
[13] Rabiner L and Juang B H 1993 *Fundamentals of Speech Recognition* (Englewood Cliffs, New
Jersey: PTR Prentice-Hall, Inc.)