TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 1, February 2020, pp. 407~418
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i1.13958

Continuous speech segmentation using local adaptive thresholding technique in the blocking block area method

Roihan Auliya Ulfattah, Sukmawati Nur Endah, Retno Kusumaningrum, Satriyo Adhy
Informatics Department, Faculty of Science and Mathematics, Diponegoro University, Indonesia

Article Info

Article history:

Received Aug 23, 2019
Revised Nov 20, 2019
Accepted Dec 20, 2019

Keywords:

Blocking block area
Continuous speech
Local adaptive thresholding
Speech segmentation

ABSTRACT

Continuous speech is a form of natural human speech that is continuous without a clear boundary between words. In continuous speech recognition, a segmentation process is needed to cut the sentence at the boundary of each word. Segmentation becomes an important step because a speech can be recognized from the word segments produced by this process. The segmentation process in this study was carried out using local adaptive thresholding technique in the blocking block area method. This study aims to conduct performance comparisons for five local adaptive thresholding methods (Niblack, Sauvola, Bradley, Guanglei Xiong and Bernsen) in continuous speech segmentation to obtain the best method and optimum parameter values. Based on the results of the study, Niblack method is concluded as the best method for continuous speech segmentation in Indonesian language with the accuracy value of 95%, and the optimum parameter values for such method are window = 75 and k = 0.2.

Corresponding Author:

Sukmawati Nur Endah,
Informatics Department, Faculty of Science and Mathematics,
Diponegoro University,
Prof Soedarto St., S.H. Kampus Tembalang UNDIP, Semarang, Jawa Tengah, Indonesia.
Email: sukmawati020578@gmail.com

1. INTRODUCTION

Continuous speech recognition is a further development of isolated words recognition that recognizes words from a sentence using a machine learning algorithm [1]. Human speech is a continuous speech, a series of words composed continuously without a clear break between words. Continuous speech recognition technology is needed so that the machine can understand human speech in giving voice commands [2]. Speech recognition has been widely applied in various fields [3-6].

The implementation of continuous speech recognition consists of three major stages: pre-processing, feature extraction, and recognition [7]. Pre-processing functions to prepare speech signals so that feature extraction can be performed. One of the main process in pre-processing is the segmentation process. Segmentation process is a process of dividing continuous speech into basic units such as words, phonemes or recognizable syllables [8]. The lack of markers that indicate the initial and final limits of a word when speaking increasingly complicates the process of segmentation, especially when speaking continuously. In contrast to text that can be seen or given its segment boundaries by recognizing the space between words. The results of this segmentation will indirectly affect the results of recognition [9].

The research related to continuous speech segmentation has been conducted using several methods, audio and visual fusion for the domain of Turkish language [10], segmentation based on time-domain features and frequency-domain features applied in Bangla [8], Hybrid of time-domain features and

frequency-domain features and median filtering in Tamil [11], segmentation with dynamic thresholding and blocking block area inside Bangla [12]. Based on some of these studies, continuous speech segmentation can be done by converting speech signal representations into spectrogram images. The spectrogram image is then processed to produce word segments, one method that can be used is the blocking block area [9].

Blocking block area is the process of making word blocks from spectrogram images in the form of binary images through several stages, namely generating spectograms, performing dynamic tresholding with clustering algorithms on spectogram images to produce binary images and boundary detection. In this research, on word segmentation using dynamic thresholding in the blocking block area method with the addition of morphological operations and overlapping process which is then called improved blocking block area. This is done because it will be applied to the speech segment in Indonesian, because the blocking block area method in the study [12] has a bad result if applied in the Indonesian language domain. This may be due to Indonesian language which has many regional dialects, so that a word can have a different pattern.

Dynamic thresholding in research [12] uses a single threshold for the entire image or global threshold. Single threshold in such technique will be difficult to distinguish the background and the foreground fields in the spectrogram images with more than two regions due to varying intensities and noises in the images [13]. In such condition, some threshold values are needed for each pixel in a particular region using local adaptive thresholding technique.

In this study, continuous speech segmentation is performed using local adaptive thresholding technique to produce binary spectrogram images. This thresholding technique has been applied to binarization and image segmentation processes in several previous studies [14-17]. The binary image, results of the binarization is then processed using improved blocking block area method so that there will be word blocks based on the number of pixels for each column. Each block is a word segment that results from the segmentation process. There are several local adaptive thresholding methods including Niblack [18], Sauvola [19], Bradley [20], Guanglei Xiong [21] and Bernsen [22]. The performance of each method will be compared in this study for continuous speech segmentation.

2. RESEARCH METHOD

This research is divided into three steps, namely data collection, segmentation and testing. The following is an explanation of each step.

2.1. Data collection

The data was taken by recording four people who have different dialects in Indonesia. Each person says 20 sentences, that is:

S1	*abang bercerita sesuatu yang bagus*	S11	*maaf atas kejadian senin lalu*
S2	*bapak ibu pergi bersama adik*	S12	*makan kuning telur setengah matang*
S3	*bibi mulai terkenal sore ini*	S13	*masinis kereta berbaju biru tua*
S4	*cincin kawin dari bahan permata*	S14	*nanti siang saja kata berbahaya*
S5	*dia punya dua mobil hitam*	S15	*pabrik gula pasir ada lima*
S6	*hidup itu seperti sekotak coklat*	S16	*paman meninggal saat dulu sekali*
S7	*kamu jangan jadi judes juga*	S17	*pantun tentang pisang dan sayur*
S8	*kapan kita main bola pantai*	S18	*siapa suka anak kecil lucu*
S9	*karena keju adalah susu sapi*	S19	*sulap tepung terigu rasa roti*
S10	*kompor kredit berwarna merah muda*	S20	*tukang tipu sudah tertangkap juga*

2.2. Segmentation

There are five processes in continuous speech segmentation, namely Generate Spectrograms, Binarization, Morphological Operations, Improved Blocking Block Areas and Boundary Detection as shown in Figure 1.

2.2.1. Generate spectrogram

Generate Spectrogram converts sound signals into images of sound signal intensity that have different densities. The spectogram functions to identify and group the sound input in phonemic way. The image of the spectrogram is then converted into a grayscale image to be able to do binarization using local adaptive thresholding. The image of the spectrogram from the speech signal of "*dia punya dua mobil hitam*" is shown in Figure 2.

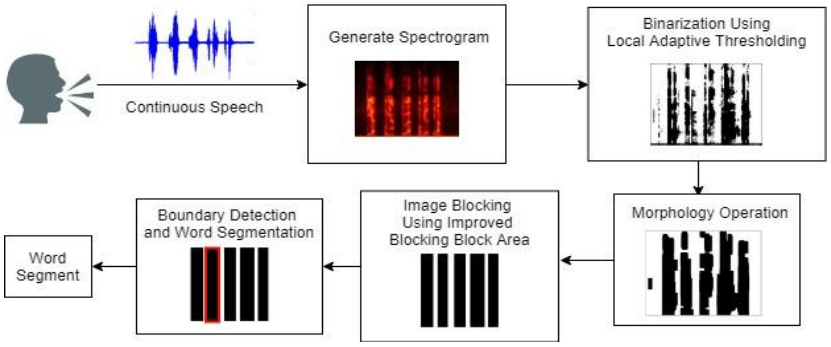

Figure 1. Process block of segmentation

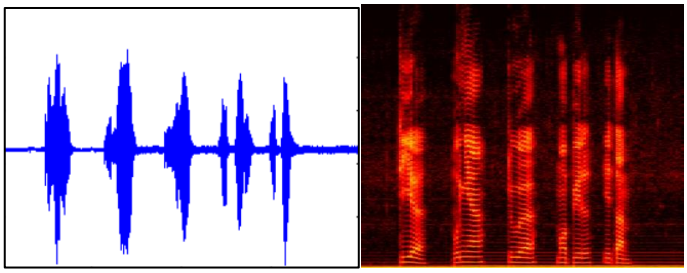

Figure 2. Speech signal and spectrogram image of "*dia punya dua mobil hitam*"

2.2.2. Binarization using local adaptive thresholding

The binary image of spectrogram is obtained through a Binarization process using local adaptive thresholding technique. This technique will produce a threshold value used to group the intensity of the input image into two values (background or foreground).

a. Niblack

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 [18]:

$$T(i,j) = m(i,j) + k \cdot \sigma(i,j) \quad (1)$$

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

The process of obtaining a binary image using Niblack is shown in Figure 3.

b. Sauvola

Sauvola determines the threshold value based on the local mean and the local standard deviation, the same as Niblack, because it is a development of Niblack method. The difference is that there is an R value in the Sauvola formula. R is the dynamic range of the standard deviation or the maximum value of the standard deviation obtained. The formula to calculate Sauvola threshold value is [19]:

$$T(i,j) = m(i,j) * \left[1 + k\left(\frac{\sigma(i,j)}{R} - 1\right)\right] \quad (2)$$

where :

m = the mean for all windows

σ = the standard deviation for all windows

k = a constant (0-1)

R = the dynamic range of the standard deviation

The process of obtaining a binary image using Sauvola is shown in Figure 4.

c. Bradley

Bradley determines the threshold value based on the local mean and the average value of brightness. The local mean is also 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. Meanwhile, the average value of brightness depends on the constant value T (in the range of 1-100). The threshold value of Bradley method can be calculated as follows [20]:

$$threshold = local\ mean * (1 - \frac{t}{100}) \quad (3)$$

the process of obtaining a binary image using Bradley is shown in Figure 5.

d. Guanglei Xiong

Guanglei Xiong determines the threshold value based on the local mean or the local median, and depends on a certain constant value in the range of 0-255. In this study, the threshold value is determined using local mean, because the average value of neighborhood for a pixel with the size of m x n is more able to represent the value of the pixel than using the local median. The threshold value of Guanglei Xiong can be calculated as follows [21]:

$$\mathrm{T} = \mathrm{mean} - \mathrm{C} \text{ atau } \mathrm{T} = \mathrm{median} - \mathrm{C} \quad (4)$$

The process of obtaining a binary image using Guanglei Xiong is shown in Figure 6.

e. Bernsen

Bernsen determines the threshold value based on the local mean, the local contrast, and the threshold value of contrast. Local contrast is the initial determinant of the threshold for two conditions, that is, if the local contrast value is less than the threshold value of contrast k, the pixel will be set to the background or the foreground depending on the global midgrey value. Whereas if the local contrast value >= the threshold value of contrast, it will be set to the background or the foreground depending on the local mean value. The threshold value of Bernsen method is calculated as follows [22]:

$$T(x, y) = 0.5(I_{\max(i,j)} + I_{\min(i,j)}) \quad (5)$$

and the provided contrast is calculated as follows:

$$C(i, j) = I_{max}(i, j) - I_{\min}(i, j) \geq \mathrm{k} \quad (6)$$

where :

$I_{\max(i,j)}$ = The maximum gray value in the local window

$I_{\min(i,j)}$ = The minimum gray value in the local window

k = The threshold value of contrast

The process of obtaining a binary image using Bernsen is shown in Figure 7.

2.2.3. Morphological operations

Morphological operations are performed to reconstruct and eliminate imperfections in the image structure the binary image [23, 24] that improve results from the segmentation process to make it more smooth.

a. Erosion

Erosion aims to reduce the edge of the object. This process matches whether there are any objects (image pixels) of the foreground that come into contact with the background, if any, the foreground value that makes a contact is changed according to the background value [25].

b. Dilation

Dilation is the opposite of erosion with the same concept [25]. This process matches whether there is a part of the element structure that comes into contact with the background when the center of the element is foreground. If there is, the background value matches the foreground value.

2.2.4. Image blocking using improved blocking block area

The improved blocking block area method aims to change the binary image of a spectrogram that has gone through morphological operations into a block image by applying the concept of overlapping columns. The method works by breaking the image into several frames then calculating the luminance value 0 and the luminance value 1 for each frame. If the luminance value 0 reaches 45% or more than the number of pixels in the frame, the frame will be colored in black. Whereas if the luminance value 1 reaches 55% or more, the frame will be colored in white. Furthermore, all of the pixel values in the frame will be changed to white if the frame is white and the next frame is white. Aside from that, all of the pixel values in the frame are changed to black.

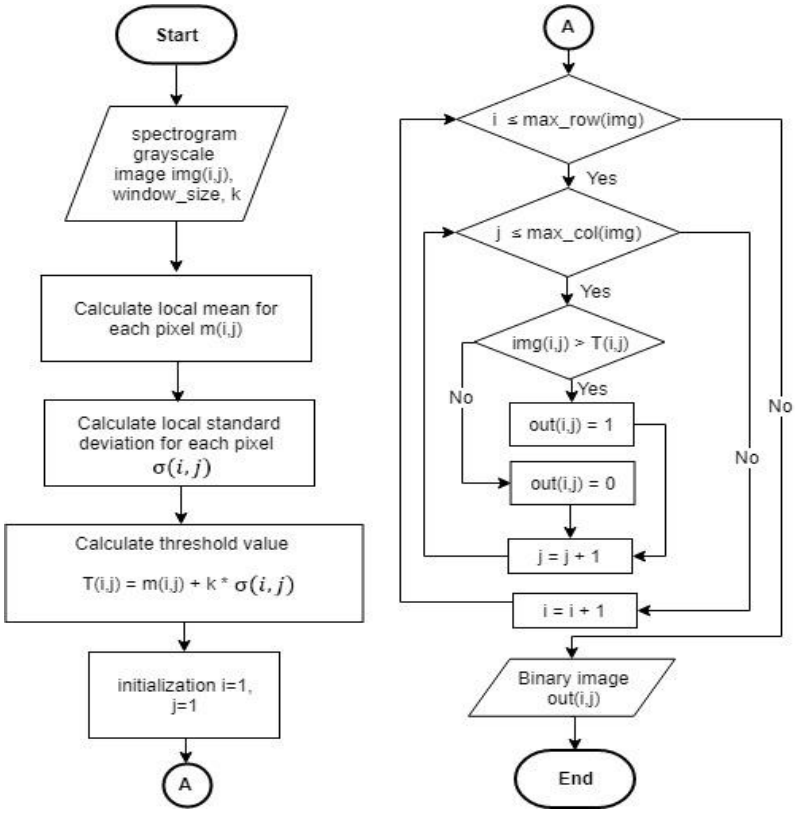

Figure 3. Flowchart to obtain binary image using niblack

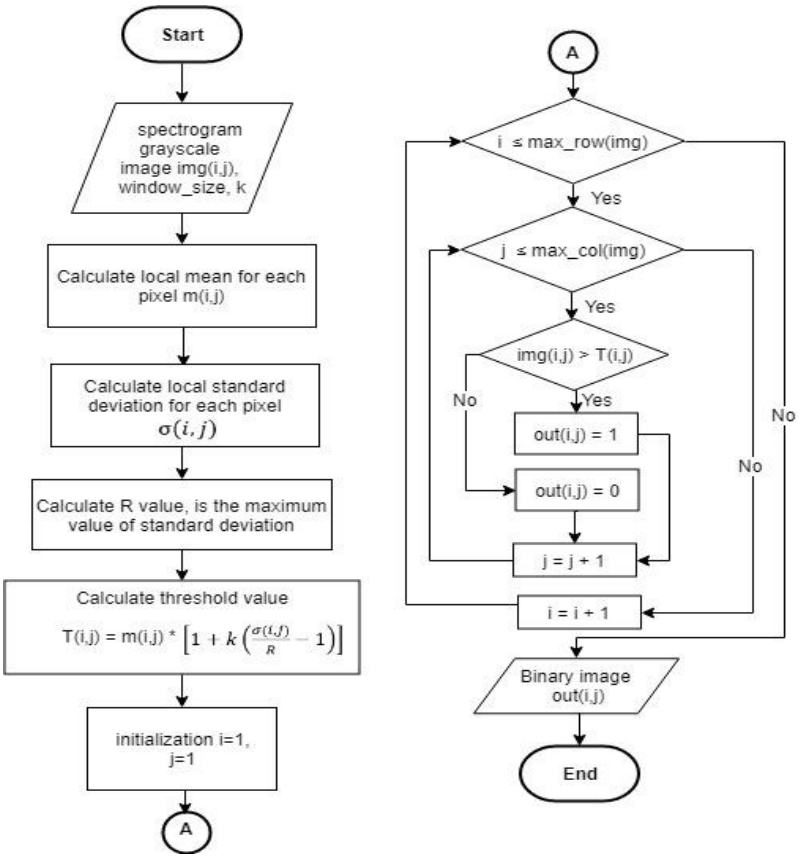

Figure 4. Flowchart to obtain binary image using sauvola

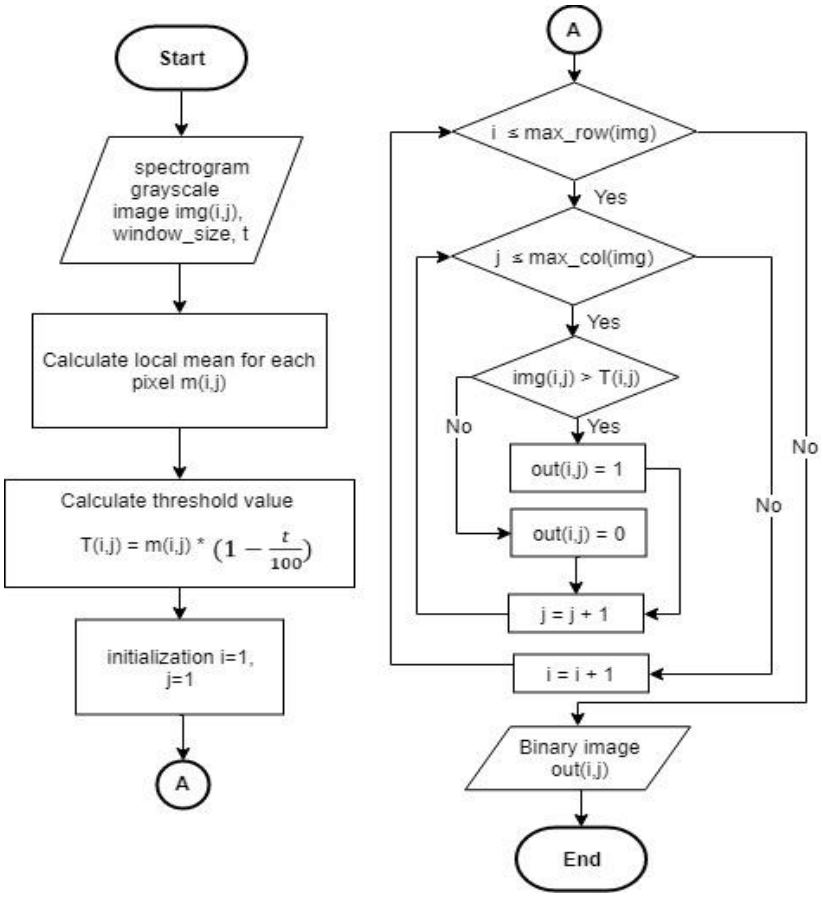

Figure 5. Flowchart to obtain binary image using bradley

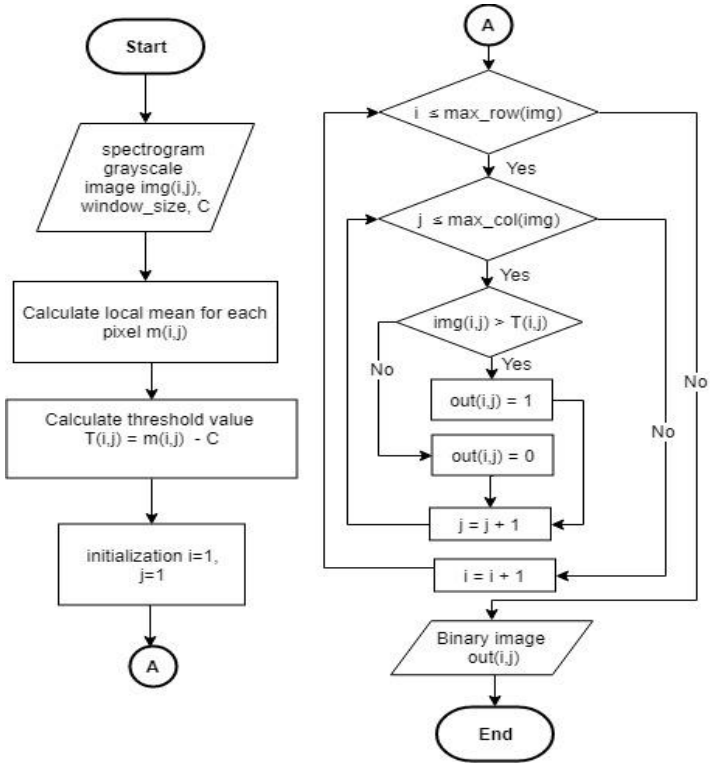

Figure 6. Flowchart to obtain binary image using guanglei xiong

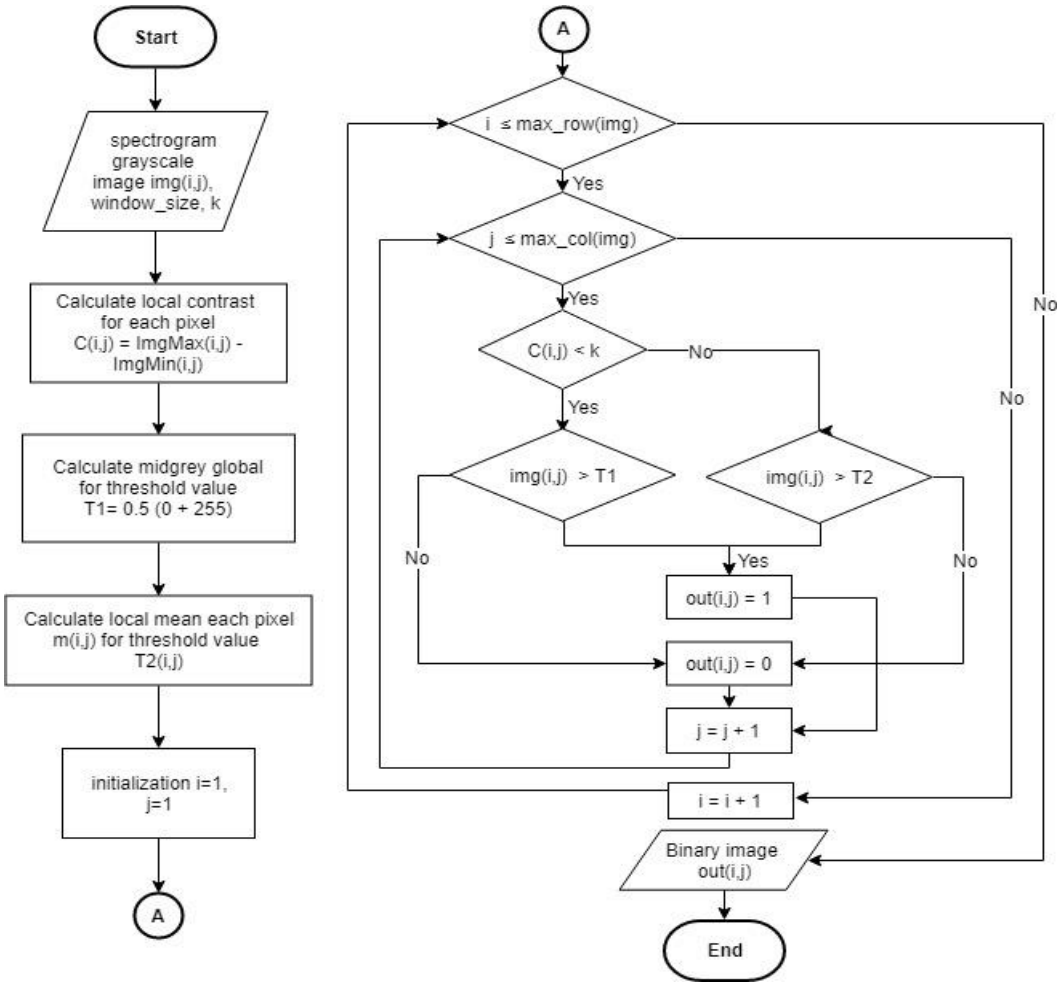

Figure 7. Flowchart to obtain binary image using bernsen

2.2.5. Boundary detection and word segmentation

The block image that has been obtained is then processed to determine the initial and the final boundaries for each block. The coordinates of the initial and the final boundaries of each block are calculated according to the overall value of the block image column. Furthermore, the results of these percentages are used as a guidance in the process of cutting the voice [12].

2.3. Testing

Segmentation testing is conducted by using five local adaptive thresholding methods and improved blocking block area method with a combination of several parameters. It has six scenarios as follows.

- Scenario 1, the segmentation uses Niblack method with increment process where the window value increase by 15 starting from window size = 15 to window size = 165.
- Scenario 2, the segmentation uses Sauvola method with a increment process where the window value increase by 15 starting from window size = 15 to window size = 165.
- Scenario 3, the segmentation uses Bradley method with a increment process where the window value increase by 15 starting from window size = 15 to window size = 165.
- Scenario 4, the segmentation uses Guanglei Xiong method with increment process where the window value increase by 15 starting from window size=15 to window size = 165.
- Scenario 5, the segmentation uses Bernsen method with increment process where the window value increase by 15 starting from window size = 15 to window size = 165.
- Scenario 6, comparing the five scenarios above.

Segmentation testing uses 80 sentences in the form of continuous speech which consists of 20 sentences spoken by 4 different people (O1, O2, O3, O4). Then calculate the accuracy in segmenting word correctly with equation as below.

$$Accuracy_{word} = \frac{\sum the\ word\ segmented\ correctly}{\sum all\ of\ the\ words} * 100\% \quad (7)$$

$$Accuracy_{sentence} = \frac{\sum the\ sentence\ with\ each\ word\ correctly\ segmented}{\sum all\ of\ the\ sentences} * 100\% \quad (8)$$

3. RESULTS AND ANALYSIS

3.1. Testing result

Figure 8 shows the accuracy of segmentation results using Niblack with the highest accuracy of 99% for word accuracy and 95% for sentence segmentation achieved when the window size is 75x75. Figure 9 shows the accuracy of segmentation results using Sauvola with the highest accuracy of 97% for word accuracy and 86% for sentence segmentation achieved when the window size is 90x90. Figure 10 shows the accuracy of segmentation results using Bradley with the highest accuracy of 88% for word accuracy and 58% for sentence segmentation achieved when the window size is 90 x 90. Figure 11 shows the accuracy of segmentation results using Guanglei Xiong with the highest accuracy of 89% for word accuracy and 60% for sentence segmentation achieved when the window size is 90x90. Figure 12 shows the accuracy of segmentation results using Bernsen with the highest accuracy of 97% for word accuracy and 86% for sentence segmentation achieved when the window size is 75x75 and 90x90.

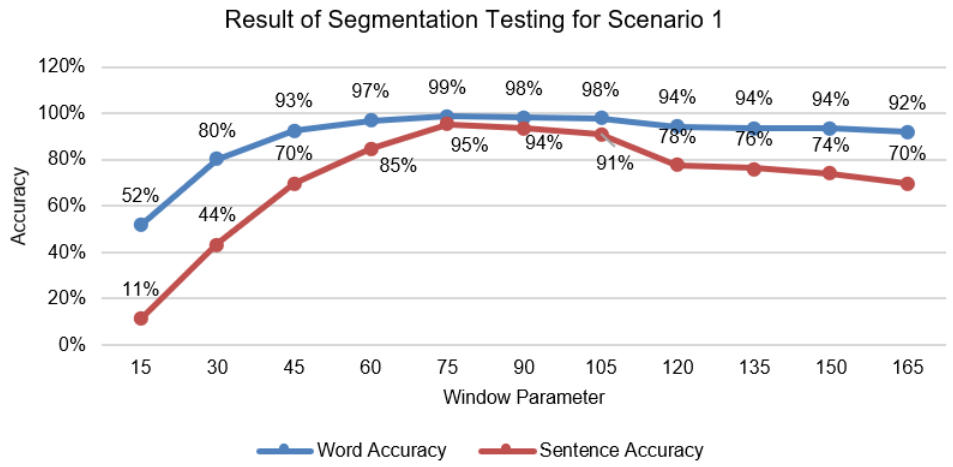

Figure 8. Graph of segmentation results for scenario 1

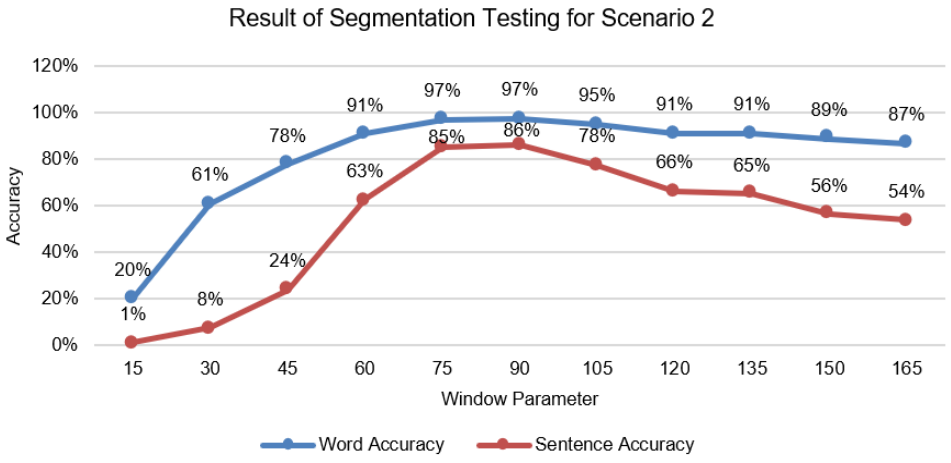

Figure 9. Graph of segmentation results for scenario 2

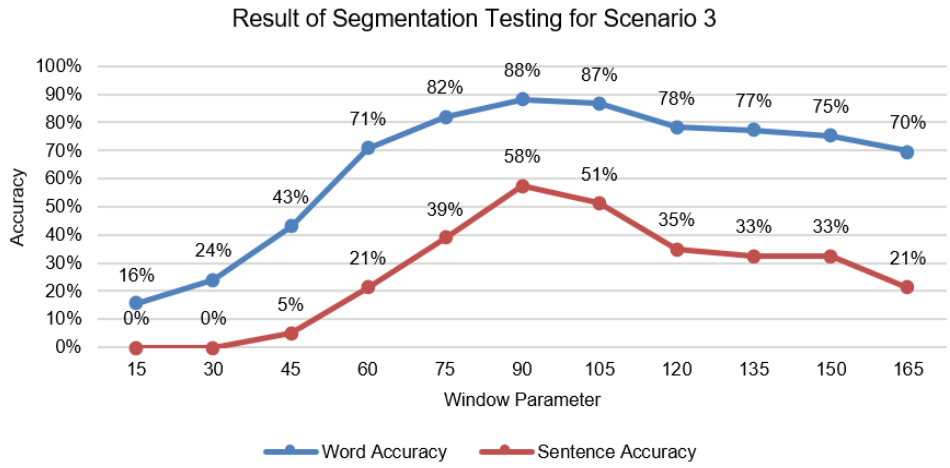

Figure 10. Graph of segmentation results for scenario 3

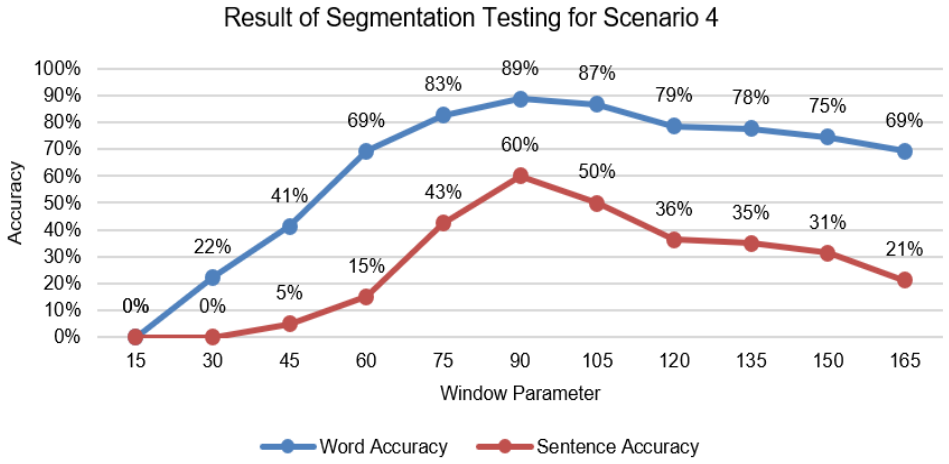

Figure 11. Graph of segmentation results for scenario 4

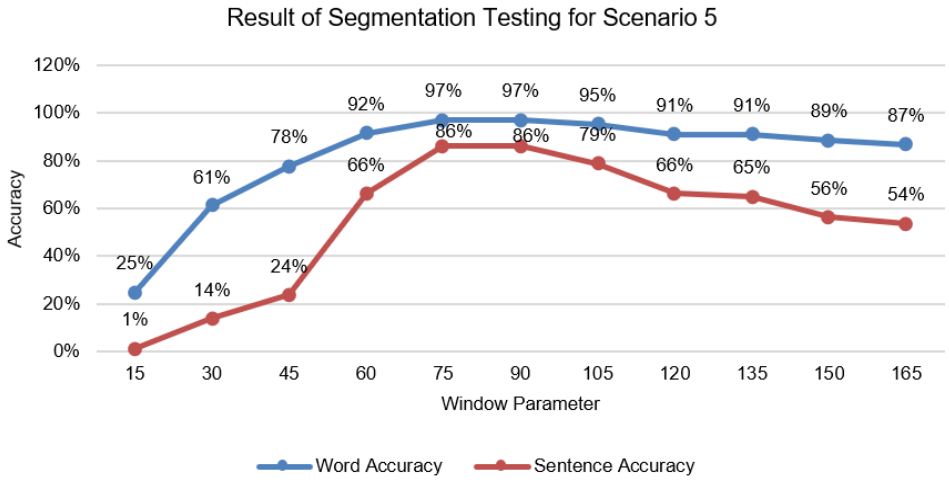

Figure 12. Graph of segmentation results for scenario 5

A comparison of the highest results of these five scenarios can be seen in Figure 13. From this figure, it can be seen that the highest accuracy of each method is achieved when the window size is 75x75 or

90x90. Above and below this window size, accuracy decreases. This can show that the optimal accuracy is in the range 60-90 or 90-105.

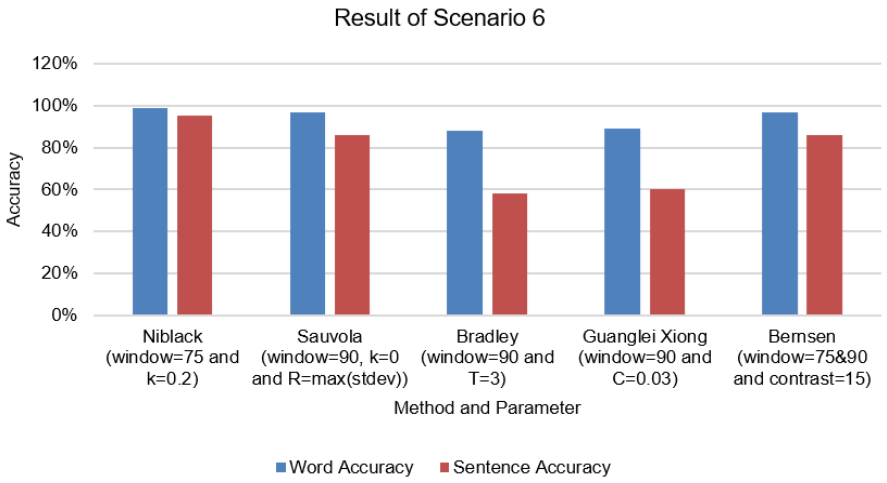

Figure 13. Graph of segmentation results for scenario 6

3.2. Analysis of segmentation result

The best result is influenced by the parameter values for Niblack method as can be seen in Figure 14 and Figure 15.

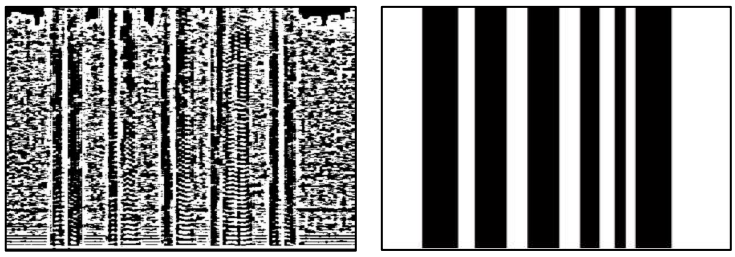

Figure 14. Result of binarization (left) and result of improved blocking block area (right) using niblack method with parameter values (window = 15 and k = 0.2)

The two images above show a comparison of the segmentation results with different window parameters that affect the segmentation results. The segmentation results with window = 75 can produce better spectrogram binary images than using window = 15, the quality of the blocking results of the spectrogram image is also indirectly influenced by the quality of the binary image.

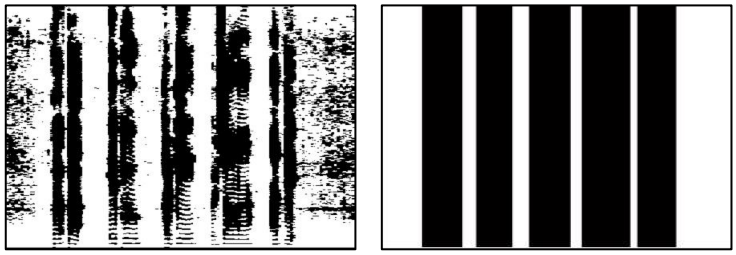

Figure 15. Result of binarization (left) and result of improved blocking block area (right) using niblack method with parameter values (window = 75 and k = 0.2)

Based on Figure 8 until Figure 12, it is shown that the window size has an effect on the accuracy of the segmentation aside from having other parameters. The accuracy of the segmentation will reach

the optimal point at a certain window range of 75 ≤ window < 105. The smaller or the larger the window size is, the smaller the accuracy will be. That is because the window size at a certain range (each method is different) can properly represent each pixel around it, so that the resulting binary image can clearly show the boundaries of each word. Besides of window size, the quality of recorded speech also influences the results of accuracy, for example accuracy produced by the first person and fourth person is always better. This is due to the effect of intonation, noise (environmental conditions, speaker sound conditions and high/low consistency of sound) at the time of recording that will be seen in the speech spectrogram image.

4. CONCLUSION

The best Local Adaptive Thresholding method for thresholding in the segmentation process in this study is Niblack method with an accuracy of 95%. The best result in segmentation process using local adaptive thresholding and improved blocking block area is influenced by parameter values in local adaptive thresholding method. The optimum parameter value is obtained using Niblack method with window = 75 and k = 0.2 which can result to a fairly good segmentation, carried out with various experiments using several parameter values.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the research funding supported by Kementrian Riset, Teknologi dan Perguruan Tinggi under the Grant of Fundamental Research for College Flagship – Contract Number 101-50/UN7.P4.3/2019.

REFERENCES

[1] N. D. Londhe and G. B. Kshirsagar, "Continuous Speech Recognition System for Chhattisgarhi," *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 365-369, 2017.
[2] F. R. Sharma and S. G. Wasson, "Speech Recognition and Synthesis Tool : Assistive Technology for Physically Disabled Persons," *International Journal Computer of Science and Telecommunication*, vol. 3, no. 4, pp. 86-91, April 2012.
[3] S. N. Endah, et al., "Integrated System Design for Broadcast Program Infringement Detection," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 13, no. 2, pp. 571-577, June 2015.
[4] S. N. Endah, et al., "Comparison of Feature Extraction MFCC and LPC in Automatic Speech Recognition for Indonesian," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 15, no. 1, pp. 292-298, March 2017.
[5] A. Jadhav and A. Patil, "A Smart Texting System For Android Mobile Users," *International Journal of Engineering Research and Applications*, vol. 2, no. 2, pp. 1126-1128, 2012.
[6] M. Pathak, et al., "Effective segmentation of sclera, iris and pupil in noisy eye images," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 17, no. 5, pp. 2346-2354, Oct 2019.
[7] A. Hasnat, et al., "Isolated and Continuous Bangla Speech Recognition : Implementation, Performance and application perspective," *Computer Science*, January 2007.
[8] M. Rahman and A. Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches," *International Journal of Advanced Computer Science and Application*, vol. 3, no. 11, pp. 131-138, Nov 2012.
[9] M. Rahman, et al., "Blocking Black Area Method for Speech Segmentation," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 2, pp. 1-6, Feb 2015.
[10] E. Akdemir and T. Ciloglu, "Bimodal automatic speech segmentation based on audio and visual information fusion," *Speech Communication*, vol. 53, no. 6, pp. 889-902, July 2011.
[11] M. Kalamani, et al., "Automatic Speech Recognition using ELM and KNN Classifiers," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, no. 4, pp. 3145-3152, April 2015.
[12] M. Rahman and A. Bhuiyan, "Dynamic thresholding on speech segmentation," *International Journal of Research in Engineering and Technology*, vol. 2, no. 9, pp. 404-411, Sep 2013.
[13] J. Rogowska, "Overview and Fundamentals of Medical Image Segmentation," 2nd, Handbook of Medical Image Processing and Analysis, Elsevier Inc., pp. 69-85, 2000.
[14] F. Shafait, et al., "Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral Images," *Proc SPIE-The International Society for Optical Engineering*, vol. 6815, January 2008.
[15] Senthilkumaran N. and Kirubakaran C., "Efficient Implementation of Niblack Thresholding for MRI Brain Image Segmentation," *Int Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 2173-2176, 2014.
[16] Senthilkumaran N. and Vaithegi S., "Image Segmentation by Using Thresholding Techniques for Medical Images," *Computer Science and Engineering: An International Journal*, vol. 6, no. 1, pp. 1-13, Feb 2016.
[17] C. Eyupoglu, "Implementation of Bernsen's Locally Adaptive Binarization Method for Gray Scale Images," *The Online Journal of Science and Technology*, vol. 7, no. 2, pp. 68-72, April 2017.
[18] N. Chaki, et al., "A Comprehensive Survey on Image Binarization Techniques," Exploring Image Binarization Techniques, New Delhi, Springer India, pp. 5-15, May 2014.

[19] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *The Journal of Pattern Recognition Society*, vol. 33, pp. 225-236, 2000.
[20] D. Bradley and G. Roth, "Adaptive Thresholding using the Integral Image," *Journal of Graphics Tools.*, vol. 12, no. 2, pp. 13-21, Jan 2007.
[21] Xiong G., "Adaptive Thresholding," *HIPR2*, [Online], Available: http://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm], 2003.
[22] T. R. Singh, et al., "A New Local Adaptive Thresholding Technique in Binarization," *IJCSI International Journal of Computer Science Issues*, vol. 8, no. 6, pp. 271-277, 2011.
[23] R. Srisha and A. M. Khan, "Morphological Operations for Image Processing : Understanding and its Applications," *NCVSComs-13 Conference Proceedings*, pp. 17-19, Dec 2013.
[24] A. M. Raid, et al., "Image Restoration Based on Morphological Operations," *International Journal of Computer Science, Engineering and Information Technology*, vol. 4, no. 3, pp. 9-21, July 2014.
[25] N. Jawas and N. Suciati, "Image Inpainting using Erosion and Dilation Operation," *International Journal of Advanced Science and Technology*, vol. 51, pp. 127-134, Feb 2013.

BIOGRAPHIES OF AUTHORS

Roihan Auliya Ulfattah, was born in Magelang, Central Java Province in Indonesia. He received the B.S degree in computer science from Universitas Diponegoro in 2019. His research interests include Artificial Intelligence and Speech Recognition.

Sukmawati Nur Endah was born in Semarang, Central Java Province in Indonesia. She received the B.S degree in mathematics from Universitas Diponegoro in 2001 and the master degree in computer science from Universitas Indonesia. She currently a lecturer in Department of Informatics, Universitas Diponegoro, Indonesia. In 2019, she is head of the research group Intelligent System. Her research interests include Artificial Intelligence and Machine Learning. S.N Endah, M.Kom is a membership of IEEE in 2017. Also, she is join the Indonesian professional organization such APTIKOM.

Retno Kusumaningrum was born in Banyumas, Central Java Province in Indonesia. She received the B.S degree in mathematics from Universitas Diponegoro in 2003 and the master and doctoral degree in computer science from Universitas Indonesia. She currently a lecturer and head of Department of Informatics, Universitas Diponegoro, Indonesia. Her research interests include Natural Language Processing and Machine Learning. Dr. Kusumaningrum is a membership of IEEE since 2015. Also, she is join the Indonesian professional organization such INACL, INAPR and APTIKOM.

Satriyo Adhy was born in Kudus, Central Java Province in Indonesia. He received the B.S degree in mathematics from Universitas Diponegoro in 2005 and and the master degree in computer science from Institut Teknologi Bandung. He currently a lecturer in Department of Informatics, Universitas Diponegoro, Indonesia. His research interests include Information System and Information Technology. He is join the Indonesian professional organization such APTIKOM and AISINDO.