Speech Signal Algorithm Conversion from Sasak Language into Sasak Script with CNN and Rule-Based Method

Arik Aranta(✉), I Gede Pasek Suta Wijaya, Fitri Bimatoro,
Gibran Satya Nugraha, Ramaditia Dwiyansaputra, and Belmiro Razak Setiawan

Department of Informatic Engineering, Univeristy of Mataram, Mataram, Indonesia
{arikaranta,gpsutawijaya,bimo,gibransn,rama}@unram.ac.id

Abstract. Sasak speech-language is can't operate in application such as google speech, Alexa assistance, and so on. It makes society need a new development technology who can solve the problem for sasak language. West Nusa Tenggara Province has more than thousand ancient manuscript writen in lontar leaves and papers, that the manuscript spread in Lombok Island and Sumbawa Island. Speech recognition technique for local language is important topic in computer science research, that it can save the culture from the island. This research should be explained how speech processing method working for conversion analog speech signal in Sasak language into Latin text. This study uses the Mel-Frequency Cepstral Coefficient (MFCC) method as a feature extraction method and Convolutional Neural Network (CNN) as a voice classification method into text, and the Rule Base method with UTF-16 which is used to provide rules on Latin letters that will be converted into text. Sasak characters. The algorithm developed is expected to be able to change 50 sound classes into Sasak letters with good accuracy results or above 88%, in changing the voice of the Sasak language into Sasak script by 90.00%.

Keywords: Speech Processing · CNN · MFCC · Sasak Script

1 Introduction

Ancient script is a regional culture that used to communicate with people in West Nusa Tenggara, long ago before massive Latin writing was used standard in Indonesia, it is currently causing the use of the Sasak script to be abandoned. Various other factors that influence the continued decline in the use of scripts, including is in terms of complexity in the writing process which it involves one factor in the reduced use of characters in the Lombok community [1]. Various studies have been carried out to preserve the use of ancient script in Indonesia, including the development of transliteration applications from Latin letters to script [2]. The other hand developing applications for changing Latin letters into characters, other research has been carried out, namely changing the image of letters into text research related to characters is very important [3]. It is important because many ancient manuscripts stored in Indonesia are written using the sasak script.

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 245–257, 2022.
https://doi.org/10.2991/978-94-6463-084-8_22

The age of this ancient manuscript is more than 50 years which contains a lot of historical information for the people of Lombok and NTB [4]. Several studies that have discussed related to ancient manuscripts have made the preservation of the sasak script. This is proof that the preservation of the Sasak script is quite massive. Human voice is one of the human senses that is used as a communication, and voice command-based systems has been increasingly used since the development of the speech to text algorithm. Voice-based in smart assistance which is the implementation of artificial intelligence developed by Google, Alxa and the like, until currently this algorithm continues to grow until the emergence of several studies that can change from text to sound [5].

In general, the use of speech processing that continues to grow makes the interaction between computers and humans easier. Behind the success of the research development, it is inseparable from the algorithms or methods, that are currently commonly used as feature-producing media and feature analysis on voice signals. This method such as Mel Frequency Cepstral Coefficient Algorithm (MFCC), which is a feature extraction algorithm that is commonly used when researchers are trying to find a distinguishing feature of each sound collection, so that the resulting features can be reprocessed to find the required information [6]. The features produced from the sound fragments cannot be directly recognized, but require a classification process in order to produce an output such as word recognition, speaker recognition, and so on, depending on the data set and the purpose of the study classification algorithms commonly used in research. Grouping data in convolution Neural Network CNN, in the CNN algorithm focus on the amount of training data and testing data in finding patterns in a data set including sound [7].

The CNN classification algorithm is commonly used in the conversion process from voice to text, which is one of the causes of the ease of the conversion process to be carried out meanwhile, rule base is an algorithm used to carry out the process of transliterating Latin letters into text. This algorithm is needed to adjust the rules possessed by character characters, which in writing characters generally do not use the character “a” in the consonants used so that if there are characters. Sound “hanacaraka” in sasak script written with the character “hncrk” this makes the process of changing from sound to text requires several additional steps to organize other text, before it is converted into script [8], so in this case it is possible for writers to be able to perform a voice change conversion process human into Latin text and the fox back into Sasak script.

Based on previous research, this research is how to do the conversion process from sound into Sasak script so that later it can help solve some of the problems found in transliterating text into sasak script, such as the pronunciation of the word “e” which has two mentions so that with this algorithm can facilitate the process of character recognition. Based on this, this research will develop an algorithm for changing hu-man voice into into sasak script text.

2 Study Literature

Several studies have continued to develop recently related to script design, including the development of speech processing algorithms, as well as developments in making Latin letter transliteration applications into certain characters. Research that has been done in the development of speech processing algorithms and script transliteration is explain on state of the art below in Table 1.

Table 1. State of The Art System.

No.	Research Title	Th	Research Area			
			CNN	MFCC	Sasak Script	Text Conversion
1	*Speech processing: MFCC based feature extraction techniques—An investigation*	2021	X	✓	X	X
2	Speech recognition using convolutional neural networks	2018	✓	✓	X	X
3	*Segment repetition based on high amplitude to enhance a speech emotion recognition*	2020	X	✓	X	X
4	*Learning media for the transliteration of Latin letters into Bima script based on android applications,*	2021	✓	✓	X	X
5	Voice Signal Conversion Algorithm into Sasak Script Using MFCC, CNN and Rule-Based Methods.	2022	✓	✓	✓	✓

Based on Table 1, no research has been found regarding the design of the Voice Signal Conversion Algorithm into Sasak Script Using the MFCC, CNN and Rule-Based Methods.

2.1 Speech Signal

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

The human voice contains several components including amplitude, time period and frequency. In the research the human voice has a unique feature called a feature, where this feature can be recognized by the feature extraction method in human analog sound processing.

2.2 Mel Frequency Cepstral Coefisien

Mel-Frequency Cepstral Coefficient (MFCC) is an algorithm used to generate a feature of the human voice, the feature is generated based on the sound pressure received by the microphone, and executed using a programming language. Using MFCC voice features can be utilized and used to perform various needs related to speech recognition [6]. The following are the stages of how the MFCC works:

1. Take the Fourier value of the voice signal sample
2. Map the power spectrum obtained from the mel peak, using triangular overlapping windows
3. Take the log value of each mel frequency
4. Take a discrete cosine transformation from the log mel power set, make it a signal
5. MFCC is the amplitude that produces the spectrum

With this MFCC algorithm, the sound collection that has been recognized in Fig. 1 will be recognized in the form of a mel frequency number line which will then be recognized by the classification algorithm. Before producing a good sound feature, an audio signal, a sound signal requires a process that becomes the stage of an MFCC process, such as windowing with a size ranging from 0.25 ms with windowing, the MFCC can produce the MFCC feature converting the audio signal into a measurement parameter generated by the mel spectrum [9] (Fig. 2).

1. Frame blocking is the process of cutting the voice signal frame, this is needed for the voice signal analysis process stage, this cutting is done up to the Nth signal to be processed by sound.
2. Pre-Emphasis is the process of suppressing a sound signal so that the results obtained can be lower, this aims to eliminate noise in the signal that can affect the accuracy of the high-level analysis process.

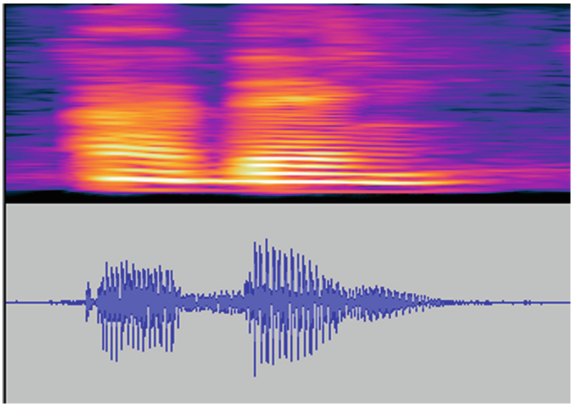


Fig. 1. Display of digital signal waves.

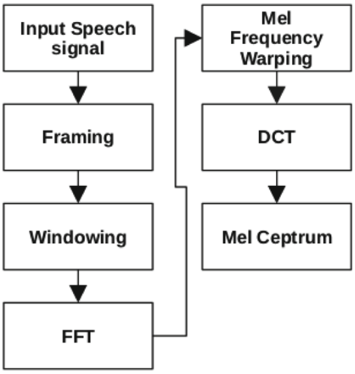


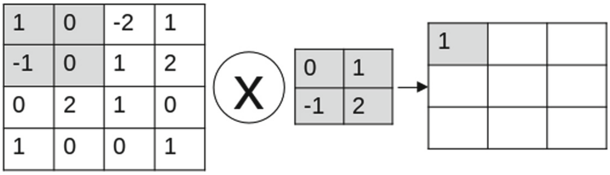
Fig. 2. MFCC Block Diagram.

3. Windowing Is a filter process on the frame that serves to eliminate the initial value and the final value of a frame, the window size at work is in the range of 25 ms, this windwing is carried out on every sound signal frame
4. FFT is a process to bring up information in a voice signal, by making a change from the voice signal that originally used the time domain to be converted into a frequency domain.
5. Mel Frequency warping is a process used to determine the size of the frequency energy of a particular band, and this process generally uses a filter bank to execute a voice signal.
6. The Discrete Cosine Transform (DCT) is a transformation similar to the discrete Fourier transform. The DCT is equivalent to a DFT that produces double the data, where a DCT computes the order of factor information in terms of the sum of the oscillating cosine features on the frequency side. The DCT module reduces the possible sign to feature coefficients with minimum dimensions.

2.3 Convolution Neural Network

Convolution Neural Network is an algorithm used to imitate human behavior in recognizing an object, in this case several studies using CNN as object recognition in the form of attractive images are used [10]. CNN is a deep learning algorithm that is very popularly used today in fields that can be solved with various CNN algorithms, ranging from voice recognition, facial recognition, and so on where the intelligence that runs on this algorithm imitates the intelligence possessed by humans. Where if a data set that has been prepared and grouped with similar data, this algorithm can recognize other inputs that have similarities with a group of data or not [11]. In the process of execution of the CNN algorithm has several stages called layers, these layers will be described as follows.

1. Kernel is a grid of numbers or a discrete value that describes how the kernel is, this kernel is taken from the data to be tested.



1	0	-2	1
-1	0	1	2
0	2	1	0
1	0	0	1

0	1
-1	2

1		

Fig. 3. Calculation of Convolutional Layer.

2. The initial format of the CNN data input is in the vector format, where the vector is the traditional input commonly found in artificial neural network algorithms. This is different from CNN which uses multi-channel as input to the CNN algorithm. With the form of operation as shown below. If there is an example of a 4 x 4 scale initialized with a random weight of a 2 x 2 kernel that executes both horizontally and vertically. Where the corresponding values are squared and then summed to get a single scalar value.

   In Fig. 3 is an illustration of the calculation process in the convolution layer, the 4 × 4 scale is multiplied by the 2 × 2 kernel and is repeated until the entire 3 × 3 kernel shows the multiplication result, carried out sequentially on each line.
3. Pooling layers is a mapping process that is used to reduce the resulting convolution calculations. New maps are generated by following the convolution operation. or it can also be called this process is the process of reducing the feature map that has been created, but still retaining most of the dominant information (features) at each step of the unification stage, with a method similar to the convolution operation, where both strides and kernels that get the size before the pooling operation are executed. Some of the pooling methods commonly used include min pooling, max pooling, global average pooling (GAP), tree pooling, gated pooling. Where the most popular pooling method is GAP pooling, maybe at some point in time CNN's performance has the potential to decrease overall, this is a weakness that CNN has. Activation function (non-linear), this non-linear tuning is a major function of all types of activation functions of various types of neural networks. Where in this function has a calculation rule by doing a weighted sum of the input neurons and their bias if the bias is found. Which can also mean that a decision-making process will occur whether to activate neurons with reference to certain inputs or not by making outputs that are in accordance with the scenario model as follows (Fig. 4).

Non-linear activation function, used when all layers are weighted or called learnable layers in the CNN architecture. Which is where the purpose of this Appendix is to explain that the input to output mapping will be processed non-linearly. More than that this layer can give CNN the ability to study more complex things.

2.4 Sasak Script

The Sasak are a tribe located in Lombok, West Nusa Tenggara, Indonesia. The culture of the Sasak people is very diverse, such as traditional clothes, songs or songs and one of them is script. Sasak script is a symbol or character that is used to write or provide

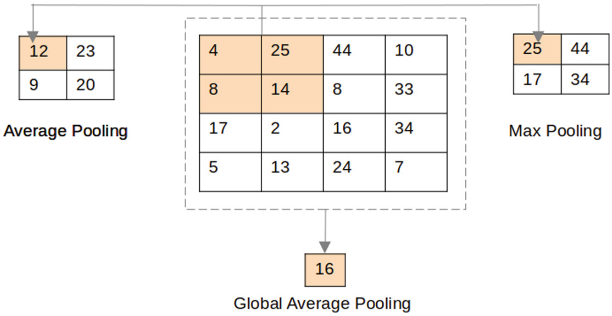


Fig. 4. Operation three on the pooling method.

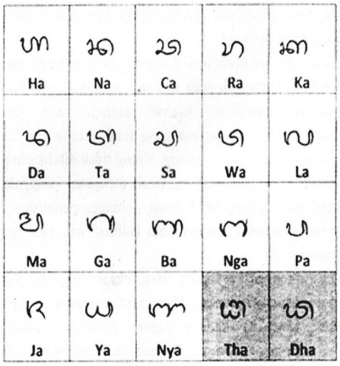

Ha
Na
Ca
Ra
Ka
Da
Ta
Sa
Wa
La
Ma
Ga
Ba
Nga
Pa
Ja
Ya
Nya
Tha
Dha

Fig. 5. Baluq Olas script in handwritten form [12].

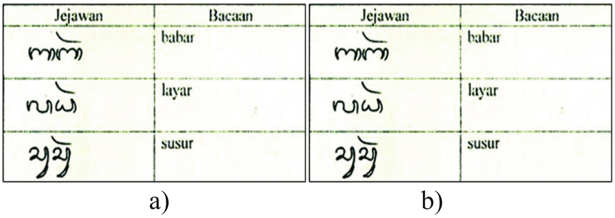
information like Latin letters that are currently used to exchange information. This script emerged since the influence of Kawi culture that entered Lombok. The Kawi manuscript entered Lombok at the same time as the wayang culture in the Sasak community [12]. The Sasak script has several components including:

Baluq Olas Script. The Baluq Olas script, the mention is based on the number of these characters being 18 characters, the original appearance of this script when written by hand is as shown in Fig. 5.

The Baluq Olas script consists of the characters ha, na, ca, ra, ka to nya, tha, dha (Fig. 6).

Face and screen are the combinations that are used to end words with the final letter “h”, and lar is the desired article to end the word “r”.

Pasang and gantungan are used to turn off the character in the middle of the kater, this is different from the swara which is used to bring out the vowel character of a consonant letter, and the hanger is to turn off the last letter of the consonant character. The appearance of the pasang and gantungan is as follows.



a)

Jejawan	Bacaan
	babar
	layar
	susur

b)

Jejawan	Bacaan
	babar
	layar
	susur

Fig. 6. **a** Sandangan Wisah, **b** Sandangan Layar [12].

2.5 Bali Simbar

Bali simbar is a font that is used to change Latin characters into traditional script characters such as Balinese script, Sasak script, and Bima script, but the change in character is only limited to changing Latin characters into traditional scripts that have not implemented the correct writing rules, knowledge of the experts is needed to compile writing from Latin letters into good and correct Sasak script [13].

3 Method Implementation

After collecting all the needs for the design of the algorithm, proceed with starting to process the data that has been obtained by making the concept of algorithm design and data processing models. From the concept that has been determined can be made as Fig. 7.

At this stage, if all research needs have been completed, the model that has been designed will execute according to the predetermined path. The process of changing voice signals into text can be seen as follows. The sound snippet owned is like the word "bale" so that it can become a Sasak character. Includes several stages such as the signal framing process which aims to eliminate unnecessary sound components, such as noise or other sounds that are not included in the core sound. After framing is complete, the next stage is windowing, which is the stage of the frame recognition process that has been obtained from the previous process where at this stage it aims to produce each frame piece before being executed in the FFT stage. In the FFT stage, the sound is originally time-based, so here the sound is changed based on frequency, this process aims to get the mel frequency which will later become the main ingredient of the DCT process, in

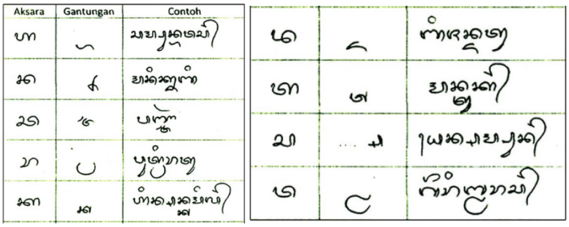

Aksara
Gantungan
Contoh

Fig. 7. Pasang and Gantungan.

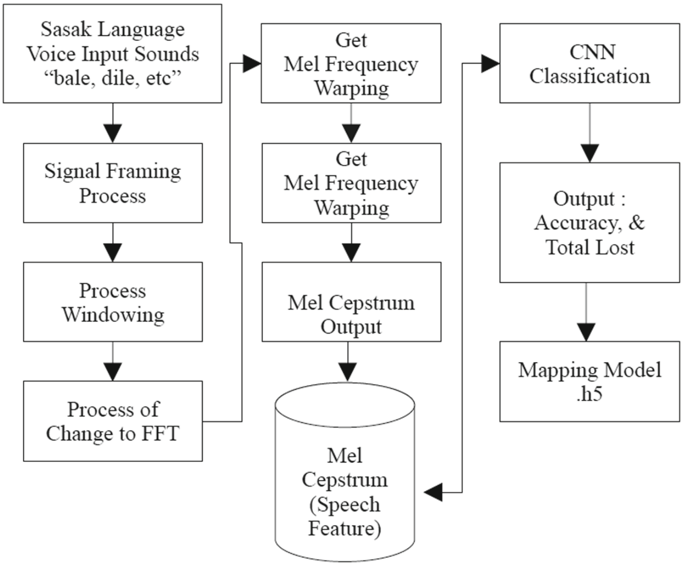

Sasak Language Voice Input Sounds "bale, dile, etc"
Signal Framing Process
Process Windowing
Process of Change to FFT
Get Mel Frequency Warping
Get Mel Frequency Warping
Mel Cepstrum Output
Mel Cepstrum (Speech Feature)
CNN Classification
Output : Accuracy, & Total Lost
Mapping Model .h5

Fig. 8. Voice Conversion Method Workflow into a mapping model.

this DCT result the mel cepstrum value is obtained and then stored in the database. The stored mel cepstrum data will then be processed using the CNN classification method which will later label the output obtained in the form of Latin text, and the Latin text obtained will be treated with a rulebase algorithm which will later be output using the Sasak script symbol as shown in the Fig. 8.

4 Result and Discussion

The testing process is carried out by evaluating the algorithm's performance on 50 times the voice input results on the sample text output data tested with training data, and testing is carried out with experimental tests carried out with real-time experiments on the system that has been designed and the last step is testing the application of the rule base to change Latin text to Sasak script.

In this study, processing 50 words of sample data with the number of data for each word is 50 sample data, using 50 sample data for each class with the aim of obtaining iteration stability from the causation value. The examples of recognized words are as shown in Table 2.

The example of feature extraction results using the MFCC method will produce 13 features from each sample consisting of 50 data from 50 data classes, where the total features obtained are 32,500 features with accuracy results as shown below. So that it can be determined what the accuracy value of the three stages of the test is for Test loss: 0.2707861065864563, test accuracy: 93.93346309661865 (Fig. 9).

In the research conducted, the results of testing voice changes into text with an accuracy of 88% with the accuracy of changing voice into text are obtained. And get

Table 2. Word Data Sample

Adeng	*Dahar*	*Inaq*	*Ngomeh*	*Saiq*
Adeq	*Dila*	*Jenu*	*Nyalean*	*Selapu*
Aiq	*Doro*	*Joman*	*Oat*	*Side*
Anjah	*Empos*	*Kadi*	*Osok*	*Sugih*
Bansu	*Endeqman*	*Kadu*	*Padu*	*Tangkong*
Bareh	*Gaur*	*Lampaq*	*Pelotan*	*Tindoq*
Bekedek	*Gaweq*	*Lekak*	*Polak*	*Uiq*
Berajah	*Harep*	*Mace*	*Priaq*	*Ulu*
Cakreh	*Hawe*	*Manuk*	*Rari*	*Wah*
Cecel	*Ima*	*Ndaq*	*Rembek*	*Wayahan*

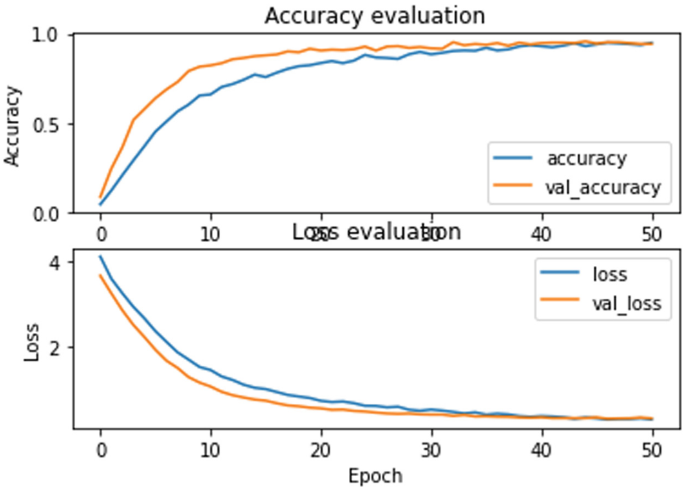


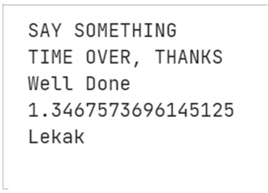
Fig. 9. Test Accuracy.

the accuracy of changing text into Sasak script with an accuracy of 90% with errors in the process of changing the use of the “q” character. After the feature extraction process is complete, the next step is to design voice calls. The scenario in the voice calling process in this study is to use the CNN method, where features that resemble the 13 characteristics produced are the features that are later suspected to be the features being searched for. The output of the intended feature of a word is as shown in Fig. 4.

After the feature extraction process is complete, the next step is to design voice calls. The scenario in the voice calling process in this study is to use the CNN method, where features that resemble the 13 characteristics produced are the features that are later suspected to be the features being searched for. The output of the intended feature of a word is as shown in Fig. 4. The MFCC Voice Features outputs derived from the MFCC

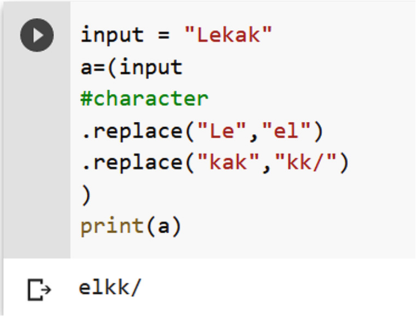
Table 3. Speech Sample Feature

1 sample sound feature	
[760.0009765625, 6.32349967956543, 6.76461124420166, 7.979455471038818, 5.272600173950195, 5.983797073364258, 5.247683525085449, 4.89335823059082, 4.646980285644531, 3.521531105041504, 5.225647449493408, 3.543940544128418, 3.2598443031311035]	[−751.2543334960938, 7.060182571411133, 10.835183143615723, 15.632367134094238, 3.979569911956787, 10.637741088867188, 4.022177219390869, 5.994346618652344, 6.763985633850098, 1.598612666130066, 10.53862476348877, 2.7283926010131836, 1.3896831274032593],



```
SAY SOMETHING
TIME OVER, THANKS
Well Done
1.3467573696145125
Lekak
```

Fig. 10. Segmentation Test



```
input = "Lekak"
a=(input
#character
.replace("Le","el")
.replace("kak","kk/")
)
print(a)
```

```
elkk/
```

Fig. 11. String replacement with rule base

process will produce outputs such as the following features. This file has addresses from 0 to 50. Where, each sound feature produced is 13 features that become the reference for a classification process, which amounts to 50 × 50 Feature (Table 3).

The stages after the feature is found and the h.5 file is generated, with this the next step is to use the librosa library, tensorflow, numpy, speech_recognition, speechpy, pydub and python display, to produce the output as follows when inputted “leak” sound (Fig. 10).

When the sound has become text, then after this stage, the next step is to take the text value and change it into the correct character arrangement. At this stage a simple step is to use string replacement as follows (Fig. 11).

When using this method, previous research has encountered several obstacles, namely, there is a collaboration that must be separated with the same characters, such as “ng” and “nga” are two different characters, this problem can be solved by changing the letters to UTF-16, on the character “l” will be \u006c do this on all characters and do a mix so as to produce a richer collaboration.

Results of Changes Using Bali Simbar, from the character “elkk/” using the Balinese Simbar font.

5 Conclusion

The research conducted, the results of testing voice changes into text with an accuracy of 88% with the accuracy of changing voice into text are obtained. And get the accuracy of changing text into Sasak script with an accuracy of 90% with errors in the process of changing the use of the “q” character.

Acknowledgment. Acknowledgments can be given to LPPM, University of Mataram, which has supported this research and to related parties who have played a role in helping to complete this research.

References

1. H. Ismi, A. Asrin, and A. Widodo.: Analysis of the Use of the Sasak Script in the Daily Life of the West Lombok Society in the Era of Globalization, AL MA’ARIEF J. Educator. Sauce. and Culture, vol. 2, no. 2, pp. 65–71, doi:https://doi.org/10.35905/almaarief.v2i2.1830 (2020).
2. A. Aranta et al., “Learning media for the transliteration of Latin letters into Bima script based on android applications,” J. Educ. Learn., vol. 15, no. 2, pp. 275–282, doi:https://doi.org/10.11591/edulearn.v15i2.19013 (2021).
3. A. A. S. M. K. Maharani and F. Bimantoro.: Recognition of Sasak Script Handwriting Patterns Using Linear Discriminant Analysis Methods and Backpropagation Types of Artificial Neural Networks, J. Teknol. Information, Computers and Apps. (JIKA), vol. 2, no. 2, pp. 237–247, doi:https://doi.org/10.29303/jtika.v2i2.105 (2020).
4. M. T. Anwar, H. Husain, and N. N. Jaya.: Digital-Based Preservation of Lombok Sasak Ancient Manuscripts and Websites, J. Teknol. inf. and Computing Science., vol. 5, no. 4, p. 445, doi:https://doi.org/10.25126/jtiik.201854787 (2018).
5. S. S. Gantayat.: Development of TTS engine for Indian accent using modified hmm algorithm, Int. J. Informatics Vis., vol. 2, no. 2, pp. 92–95, doi:https://doi.org/10.30630/joiv.2.2.112 (2018).
6. D. Prabakaran and S. Sriuppili.: Speech processing: MFCC based feature extraction techniques - An investigation, J. Phys. conf. Ser., vol. 1717, no. 1, doi:https://doi.org/10.1088/1742-6596/1717/1/012009 (2021).
7. Nagajyothi and P. Siddaiah.: Speech recognition using convolutional neural networks Int. J. Eng. Technol., vol. 7, no. 4.6 Special Issue 6, pp. 133–137, doi: https://doi.org/10.14419/ijet.v7i4.6.20449 (2018).
8. P. W. Pratama et al.: Design of Latin Script Transliteration Application into Sasak Script Using Android-Based Rule Based Algorithm,” vol. 3, no. 2, pp. 232–243 (2021).
9. B. A. Prayitno and S. Suyanto.: Segment repetition based on high amplitude to enhance a speech emotion recognition, Procedia Comput. Sci., vol. 157, pp. 420–426, doi:https://doi.org/10.1016/j.procs.2019.08.234 (2019).

10. H. Hakim and A. Fadhil.: Survey: Convolution Neural networks in Object Detection, J. Phys. conf. Ser., vol. 1804, No. 1, doi:https://doi.org/10.1088/1742-6596/1804/1/012095 (2021).
11. L. Alzubaidi et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, vol. 8, no. 1. Springer International Publishing (2021).
12. K. P. NTB, Mulok Sasak Class XII, I. West Nusa Tenggara: NTB, KEMENDIKBUD PEMPROV (2018).
13. Made Suatjana, Babad Bali - Bali Simbar Aksara Bali Bali Galang Foundation. https://www.babadbali.com/aksarabali/balisimbar.htm, 2003.