Indonesian Alphabet Speech Recognition for Early Literacy using Convolutional Neural Network Approach

Duman Care Khrisne[1] and Theresia Hendrawati[2]

[1]Department of Electrical Engineering
Faculty of Engineering, Udayana University
Badung - Bali, Indonesia
duman@unud.ac.id

[2]Informatics Department
STMIK STIKOM Indonesia
Denpasar - Bali, Indonesia
theresia.hendrawati@stiki-indonesia.ac.id

Abstract - Games are considered capable of being used as a learning medium that can help teachers to teach children how to pronounce the Indonesian alphabet in early literacy, we try to build one aspect of the game in this study. The approach we use is a speech recognition approach that uses the convolutional neural network method. The results of this study indicate that CNN can recognize speech, with input data is in the form of sound. We use the MFCC feature vector sound feature to make a 3-dimensional matrix of input sound into CNN input. We also use the Sequential CNN architecture made from a simple 10 layer neural network, which produces a model with a small size, approximately only about 6 MB, with high accuracy (84%) and an F-Measure of 0.91.

***Index Terms*— CNN, Indonesian Alphabet, MFCC, Speech Recognition.**

I. INTRODUCTION

Indonesian (*Bahasa Indonesia*) is a unifying language as well as a national language that is used by everyone in Indonesia. *Bahasa Indonesia* is a phonetic language, which is a language with a direct relationship between spelling and pronunciation. One can look at a written Indonesian word and know how to pronounce it, or we can hear an Indonesian word and know how to spell it [1]. Early literacy is a learning activity that aims to develop children's literacy [2]. This activity can be carried out since the child is four years old [3], by carrying out various approaches, such as the introduction of the sounds of letters and symbols, grammar, and vocabulary [4]. The earliest literacy for children is to learn to recognize letters or symbols, and try to pronounce it, this is a very good start as a basis for children's learning in subsequent reading. Although early literacy has an important role in children's language development and reading ability. Some problems must be resolved to make this process, can be carried out properly. As we know, childhood is a time to play, so the learning process in early literacy should not become a burden for children.

One approach that can be used to help the process of early literacy is to build learning media that can help children feel that the learning process being done is a fun activity or game. To build a game that can help children recognize and pronounce the alphabet symbols. A computer game system is needed with two capabilities. First the game can help the process of displaying visuals to recognize symbols. The second ability, the game must be able to help the process of learning the pronunciation. For the first ability, utilizing computer visuals is sufficient to solve symbol recognition problems. Coupled with the presence of multimedia devices such as speakers, it is very likely the computer game system will be able to complete this kind of task. But it is very different from the pronunciation ability, the game must be able to process user input in the form of sound and perform speech recognition.

Therefore in this study we try to make a game, as a learning medium that can help teachers to teach children how to pronounce the Indonesian alphabet. The approach we use is a speech recognition approach that uses the convolutional neural network method.

II. RELATED WORKS

Previous studies have shown that several approaches in early literacy to solve the problems mentioned above. Fitta et al. [5] emphasized that teachers are the most important

Corresponding author Duman Care Khrisne – duman@unud.ac.id

aspects, so training is needed for teachers to master how to introduce the alphabet to children by utilizing learning media which they call the *alphabet book smart kids*.

Research [2] uses phoneme recognition techniques to help learning English. The introduction of the phoneme is part of the phonic method used in learning English, especially to improve reading skills in children in the early childhood education environment. From this study we got information that the phonic method can be done as a beginning of learning to read for early childhood by introducing symbols and the sound of letters.

Research in [7] and [8] attempted to design an augmented reality (AR) educational game application with the aim of being both educative and fun in helping to teach the alphabet in Indonesian. The results of this study are AR applications that display letter characters and examples of the use of letters on objects in the environment around children. Educational games are indeed a promising way for children's learning, this is proven by research [8] that also uses interactive puzzle educational games as learning media to teach the alphabet.

In this study we will also use games as a learning medium for children to learn the pronunciation of the alphabet. But pronunciation is an activity that requires hands-on training, so we will try to approach speech recognition in the educational game that we have designed.

Mel Frequency Cepstral Coefficients (MFCC) is one of the voice feature extraction techniques that are often used as in [9]–[12], MFCC is used to distinguish one sound from other sounds. By utilizing classification techniques, we can classify MFCC features from voice input to a class that we have specified. However, the many variations of sound can cause the classification process to look for non-linear correlations. To solve non-linear correlations problem many researchers try to use machine learning techniques [9]–[16], and what is more promising is the classification using deep learning techniques. With this in mind, we decided to try the deep learning approach (Convolutional Neural Network) on the speech recognition module in our educational game application.

III. Proposed Approach

The speech recognition approach that we used in this study is a Convolutional Neural Network (CNN), CNN is a form of artificial neural network that has a 3-dimensional input type. Because the sound input that has been extracted using the MFCC feature only leaves a 1-dimensional vector shape, we have to make a few changes to our feature vector as input for CNN.

We use MFCC feature vectors with size 11, we make it constant because CNN cannot process vectors of varying sizes although MFCC vectors might vary in size for different audio input. For that we have to make an MFCC feature vector that uniforms in size. If after the MFCC process is obtained more than 11 elements in the extracted feature vector then the excess will be removed, whereas if less than 11 features will be padding by filling in the remaining vectors with the number 0.

The sound input will be sampled with a number of 20 samples per sound, and we will extract each sample with MFCC and give us 11 features, so now we will have a two-dimensional matrix that represents the number of features and sound sampling. This matrix will have a size of 20 × 11. By having a 2-dimensional matrix, to make it a 3-dimensional matrix we only need to change it in the context of the program, because a 2-dimensional matrix can be considered a 3-dimensional matrix with a depth of 1 (number of sound channel that we use). Fig. 1, shows how the process of embedding sound into vector shapes we did in this study.

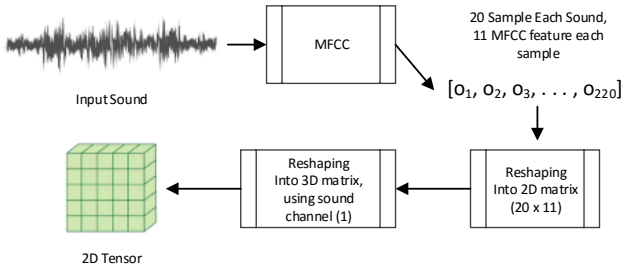

Fig. 1. Sound embedding to make an input tensor for CNN

Keep in mind, this method can be done because the type of speech recognition we expect in the game is a one-word (one pronunciation of the alphabet) recognition type. This allows us to simplify our input, without thinking about embedding other words in the input sound. The design of the game and the use of speech recognition in the game are shown in Figure 2

To build a speech recognition model (CNN) in this study, we had to go through 2 stages before getting the model we wanted. The first stage is building training data and the second stage is conducting training to build our CNN model. To go through these two stages we use 104 sound files, with .wav file types, with a maximum duration of 1 second with a sample rate of 44100 and a channel of 1. 104 files consist of 4 files (data) for each of the 26 classes (letters of the alphabet). We will use 3 data from each class for training data and the remaining 1 data will be used as test data. So from 104 data, 75% (78 data) will be used as training data and 25% (26 data) will be used as test data. We obtained data from [17], [18] and some additional data was taken by recording the measured respondents.

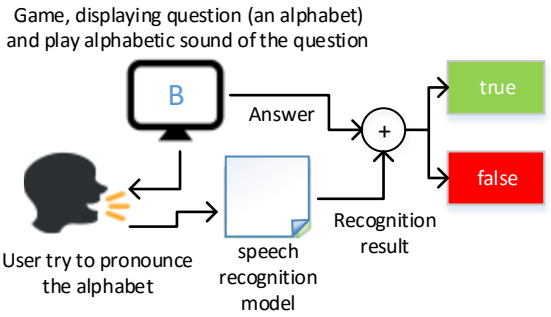

Fig. 2. Game and speech recognition process

IV. RESULT

Because using CNN we have to consider the model size and the resulting classification speed. We did some testing and got an architecture that we felt was simple enough to reduce the time needed for the classification process while maintaining accuracy. We use sequential CNN architecture with 10 layer, the architecture can be seen in Table I.

TABLE I
CNN ARCHITECTURE

No	Layer Type	Number of Filter	Kernel size / Pool Size	Activation
1	CONV	32	Kernel = 2 × 2	RELU
2	CONV	48	Kernel = 2 × 2	RELU
3	CONV	48	Kernel = 2 × 2	RELU
4	POOL	Max Pooling	Pool = 2 × 2	
5	DROPOUT	25%		FLATTEN
6	DENSE	128		RELU
7	DROPOUT	25%		
8	DENSE	64		RELU
9	DROPOUT	40%		
10	DENSE	26		SOFTMAX

We conduct training with as many as 500 epochs of the training sound dataset that we have separated with sound validation data. We get pretty good accuracy with training accuracy reaching 100% and training losses close to 0. Then we use the model generated from the training results to validate the 26 validation sound data. The training accuracy versus loss, plot result can be seen in Fig. 3. Form Fig. 3. We get information that before the 250th epoch the accuracy value of the training is already close to 100% but the loss value is still not convergent, but after the 300th epoch the accuracy and loss values begin to converge, and the model becomes stable in doing classification. The model itself only cost around 6 MB in size and can use to predict 3 second length input sound with just around 1 second waiting time.

Fig. 3. Training plot for speech recognition model using CNN

After getting the model, we conducted the validation process using 26 validation data that had been prepared. Because the classification is expected to classify into 26 classes we will calculate the accuracy, precision, and recall of the model we produce using this validation data. For this reason, we built the classification confusion matrix from this model, the classification confusion matrix can be seen in Fig. 4.

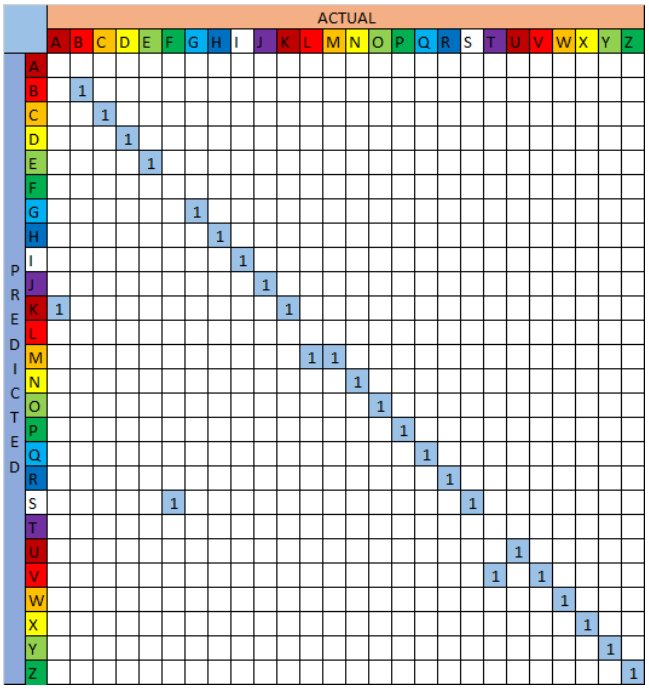

Fig. 4. Confusion matrix of validation data classification

From the confusion matrix (Fig.4.) we can calculate the model's accuracy as follows:

$$Accuracy = \frac{correctly\ classified\ data}{number\ of\ data\ to\ be\ classify} \times 100\%$$

$$Accuracy = \frac{22}{26} \times 100\% = 84\%$$

We also want to know, when it predicts something, how often is the prediction states correct class, so we need to calculate precision of the model as follow:

$$Precision = \frac{True\ Positive}{(T.Positive + F.Positive)} \times 100\%$$

$$Precision = \frac{22}{26} \times 100\% = 84\%$$

And we cannot forget about the imbalanced classification problems, so we need the F Measure, before that, we need the recall value of the classification model, we calculate it as follow:

$$Recall = \frac{True\ Positive}{(T.Positive + F.Negative)} \times 100\%$$

$$Recall = \frac{22}{22} \times 100\% = 100\%$$

After we get the recall and precision value, we can calculate the F Measure of classification model, we use the F1-Measure to get balance measurement of recall and precision using equation as follow:

$$F - Measure = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

$$F - Measure = \frac{2 \times (0.84 \times 1)}{(0.84 + 1)} = 0.91$$

By getting the F-Measure from the model we know the classification performance produced by the model is very good (0.91 out of a maximum of 1.00). Then we can conclude the model that we built to classify the alphabet pronunciation is good enough and can be implemented in the game.

V. CONCLUSION

This study shows that CNN is capable of recognizing speech, with data in the form of sound. The data must go through several stages to become a tensor, in general sound data and tensor is alike and can be changed without losing the meaning of the sound. We also found a sequential CNN architecture that can give small size model around 6 MB with accuracy of 84% which is high, with an additional F-Measure value reaching 0.91 to ensure that voice recognition can be done using this model.

Although this approach and model has a high success in recognizing sounds, but it needs to be recalled, this model is only a model used to recognize a word, so this model cannot be used to recognize sentences. This opens up opportunities for further research that the CNN approach can also be used to do speech recognition, perhaps some changes must be made so that CNN can recognize longer and more complex sounds.

REFERENCES

[1] Y. Karlina, A. Rahman, and R. Chowdhury, "Designing Phonetic Alphabet for Bahasa Indonesia (PABI) for the teaching of intelligible English pronunciation in Indonesia," *Indones. J. Appl. Linguist.*, vol. 9, no. 3, pp. 724–732, 2020, doi: 10.17509/ijal.v9i3.23223.

[2] S. M. Westhisi, "Metode Fonik dalam Pembelajaran Membaca Permulaan Bahasa Inggris Anak Usia Dini," *J. Tunas Siliwangi*, vol. 5, no. 1, pp. 23–37, 2019.

[3] S. P. Suggate, E. A. Schaughency, and E. Reese, "Children learning to read later catch up to children reading earlier," *Early Child. Res. Q.*, vol. 28, no. 1, pp. 33–48, 2013, doi: 10.1016/j.ecresq.2012.04.004.

[4] D. C. Castro, M. M. Páez, D. K. Dickinson, and E. Frede, "Promoting Language and Literacy in Young Dual Language Learners: Research, Practice, and Policy," *Child Dev. Perspect.*, vol. 5, no. 1, pp. 15–21, 2011, doi: 10.1111/j.1750-8606.2010.00142.x.

[5] A. Nurrahman, T. Wahyuni, and N. Thoyyibah, "Pelatihan Pengenalan Alfabet bagi Guru PAUD di Samigaluh Kulonprogo Alphabet Introductory Training for Early Child Teachers in Samigaluh Kulonprogo," *J. Millenn. Community*, vol. 2, no. 1, pp. 33–37, 2020.

[6] Y. A. Makambahe, D. R. Kaparang, A. Mewengkang, P. Teknologi, and U. N. Manado, "PENGEMBANGAN GAME EDUKASI PENGENALAN HURUF BERBASIS AUGMENTED," vol. 6, no. 3, 2018.

[7] E. Sinduningrum, R. Rosalina, and A. M. Hilda, "Pemanfaatan Teknologi Augmented Reality Untuk Media Pengenalan Huruf Alfabet Pada Anak Usia Dini," *J. SOLMA*, vol. 8, no. 1, p. 142, 2019, doi: 10.29405/solma.v8i1.3151.

[8] A. Syukur and A. Fitra, "Game Interaksi Pengenalan Huruf dan Perangkaian Kata," in *Seminar Nasional Teknologi dan Multimedia 2017*, 2017, no. 2017, pp. 7–12.

[9] Harvianto, L. Ashianti, Jupiter, and S. Junaedi, "Analysis and Voice Recognition in Indonesian Language using MFCC and SVM Method," no. 2011, pp. 131–139, 1978.

[10] G. Kour, M. Tech, S. Rbiebt, N. Kharar, and A. Mehan, "Music Genre Classification using MFCC, SVM and BPNN," *Int. J. Comput. Appl.*, vol. 112, no. 6, pp. 975–8887, 2015, [Online]. Available: www.ijcaonline.org.

[11] D. C. Krishne and T. Hendrawati, "Klasifikasi Musik Latar untuk Aktivitas Balita menggunakan Metode MFCC, LVQ dan DTW," *J. S@cies*, vol. 7, no. 1, pp. 42–46, 2016.

[12] I. D. G. Budi Dharma Prabhawa, D. Care Khrisne, and M. Sudarma, "Rancang Bangun Aplikasi Pengenalan Pupuh Bali Menggunakan Metode Mel Frequency Cepstral Coefficients," *J. Comput. Sci. Informatics Eng.*, vol. 3, no. 1, p. 75, 2019, doi: 10.29303/jcosine.v3i1.237.

[13] P. A. Wicaksana, I. M. Sudarma, and D. C. Khrisne, "PENGENALAN POLA MOTIF KAIN TENUN GRINGSING MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK DENGAN MODEL ARSITEKTUR ALEXNET," *J. Spektrum*, vol. 6, no. 3, pp. 159–168, 2019.

[14] T. Hendrawati, I. N. Sukajaya, and K. Y. E. Aryanto, "Automatic Image Annotation using Minimum Barrier Salient Object Detection and Random Forest," in *2018 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2018, pp. 305–310, doi: 10.1109/ISITIA.2018.8711110.

[15] D. C. Khrisne and I. M. A. Suyadnya, "Indonesian Herbs and Spices Recognition using Smaller VGGNet-like Network," in *2018 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS)*, 2018, pp. 221–224, doi: 10.1109/ICSGTEIS.2018.8709135.

[16] I. M. Wismadi, D. C. Khrisne, and I. M. A. Suyadnya, "Detecting the Ripeness of Harvest-Ready Dragon Fruit using Smaller VGGNet-Like Network," *J. Electr. Electron. Informatics*, vol. 3, no. 2, p. 35, 2020, doi: 10.24843/jeei.2019.v03.i02.p01.

[17] TheBelajarIndonesia, "BelajarIndonesia: INDONESIAN ALPHABET," *YouTube*, 2012. https://www.youtube.com/watch?v=kuzq4VKqXJM.

[18] TLIndonesian, "Indonesian Alphabet Pronunciation Guide," *YouTube*, 2014. https://www.youtube.com/watch?v=RQ4M5-v6_JQ&feature=youtu.be.