ISSN 2745-3502
Volume 3 No. 1 – Oktober 2022
Indonesian Applied Physics Letters

Speech Synthesis Based on EEG Signal for Speech Impaired Patients by Using bLSTM Recurrent Neural Network

Abdufattah Yurianta[1, a)], Anaqi Syaddad Ihsan[1, b)], Arijal Ibnu Jati[1, c)], Osmalina Nur Rahma[1, d)] and Aji Sapta Pramulen[2, b)]

[1]*Biomedical Engineering, Physics Department, Faculty of Sciences and Technology, Universitas Airlangga, Surabaya, East Java, Indonesia, 60115*

[2]*Multimedia Broadcasting, Creative Multimedia Engineering Department, Electronic Engineering Polytechnic Institute of Surabaya, Surabaya, East Java, Indonesia, 60111*

a)Corresponding author: abdufattah.yurianta-2018@fst.unair.ac.id
b)anaqi.syaddad.ihsan-2018@fst.unair.ac.id
c)arijal.ibnu.jati-2018@fst.unair.ac.id
d)osmalina.n.rahma@fst.unair.ac.id
e)aji@pens.ac.id

Abstract. Speech impairment The disability rate in Indonesia is still relatively high and is one of the main health problems which reaches 30.38 million people or 14.2% of the Indonesian population. One of these types of disabilities is speech impairment. There are several possible causes for speech impairment, including the focal disturbance. This situation occurs because of disturbances in the vocal cords caused by injuries due to accidents and other conditions, such as throat cancer, which of course will reduce the productivity of the sufferer. Sign language can be used to communicate, but it still has limitations for normal individuals. In addition, speech synthesis using brain computer interface (BCI) based on electrocorticography (ECoG) has been developed. However, this method still has a weakness, namely invasive and allows the emergence of large enough scar tissue, so that it can reduce the quality of brain biopotential to be recorded. Therefore, a non-invasive EEG-based speech synthesis method was initiated. This method uses bLSTM as one of the components of the RNN model, so that it can construct syllables into words. This system consists of datasets, data filter programs, data segmentation programs, feature extraction programs, ANN and RNN deep learning model training programs, and text-to-speech programs. ANN and RNN form a 2-level deep learning. The testing accuracy and accuracy of the ANN are 26.04% and 20.83%, while the accuracy of the RNN is 81.25%. To improve these results, in the future, researchers can improve the data collection process and increase the number of the data, use the correct extraction feature, and compare several machine learning architectures, to produce optimal accuracy.

Keyword. Speech Impairment, EEG, Speech Synthesis, ANN, bLSTM

INTRODUCTION

Speech impairment is a condition when a person experiences abnormalities, both in language pronunciation (articulation) and voice, from his normal state, causing difficulties in verbal communication with his environment. Therefore, speech impairment is classified as a disability disorder. The disability rate in Indonesia is still relatively high and is one of the main health problems which reaches 30.38 million people or 14.2% of the Indonesian population[1]. People with disabilities need to get attention from the government, given the large number and the existence of citizen rights for everyone.

According to the cause, there are several types of speech impairment, which can be suffered by adult individuals, one of which is vocal disturbance. This type of speech impairment suffers from disorders of the vocal

cords caused by injuries due to accidents and other conditions, such as throat cancer, which of course will add to the obstacles in the sufferer. This condition will certainly provide obstacles for sufferers in several aspects. The physiological, psychological, and health barriers experienced by speech impaired people will change the previously effective state, thereby reducing the productivity level of the sufferer[4].

In order to increase productivity again, speech impaired people can use non-verbal communication or sign language to communicate with other individuals. Sign language can be done with hand movements, combinations of hand shapes, to facial expressions to convey communication information. However, due to the need for special skills to use sign language, not all levels of society can use it. This of course will create a gap or barrier between people who are mute and non-speaking individuals in communicating[5]. Therefore, we need a technology that can produce sound directly from the body signal stimulus of a speech impaired person that will be issued through an audio media, and packaged in practicality and ease of meaning by anyone.

One of the sound synthesis developments that have been carried out is the use of brain computer interface (BCI) technology with the electrocorticography (ECoG) method. This method can record biosignal activity in the brain, then the bioelectric activity will function as a stimulus that will be processed to produce sound synthesis directly[2]. ECoG has the disadvantage of requiring a surgical operation to implant signal-tapping electrodes under the patient's scalp. After surgery, the possibility of scar tissue is quite large. This will have an impact on decreasing the quality of the brain's biopotential to be recorded.

Based on previous problems and inadequate technology, we are trying to develop non-invasive speech synthesis system based on EEG signals in speech impaired people that can translate neural activity into speech. This system will becomes transformative for speech impaired caused by certain accidents or diseases, thus causing vocal disturbance. We design an artificial neural network that uses entropy and other feature extraction from the recorded dataset, and applies the bLSTM model to form speech synthesis. The selection of bLSTM as one of the model components was carried out because it was proven to be used to produce speech synthesis with a large number of sentences[3]. This system is in the form of software with data acquisition with EEG connected to a PC. The system made is able to classify four syllables (a, ku, ma, kan) and still requires development in the form of adding datasets.

RESEARCH METHODOLOGY

The tools used in this research are BITalino Plugged Kit BIT and BITalino Board Kit BIT as EEG signal acquisition tools; Buff to tighten the electrode attachment to the scalp; a seam meter for measuring the position of the electrodes; markers to mark data acquisition points; smartphone for recording sound during dataset creation; as well as a personal computer (PC) to help create datasets and create software in the form of speech synthesis prototypes. The material used is the electrode for the acquisition of EEG data.

The developed system flow includes some steps, namely acquisition of EEG signal through BITalino; Filter and apply segmentation method to obtain EEG signal part for particular syllable; Extract features and establish syllables classification model using artificial neural network (ANN); Build syllables prediction model using bidirectional Long Short-Term Memory (bLSTM); and convert sentence that is constructed from syllables into sound. The system flow chart is shown in FIGURE 1.

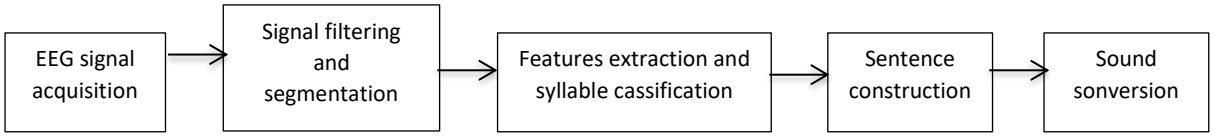


FIGURE 1. System block diagram.

The data taken for the experiment were 10 syllables, namely a, ku, ki, ta, ma, in, kan, bo, la, and yam. Each syllable is read five times in one data collection, then repeated 15 times. Each syllable contains 2 methods, namely one time visualization (imagining the pronunciation of syllables) and one time speaking (saying the syllables). The duration of retrieval of one data lasts approximately 4 minutes 10 seconds.

RESULT AND DISCUSSION

The dataset was collected by taking primary data offline at the residence of each participant. There are 6 participants who have been included in the dataset. Each participant was asked to collect data by following the health protocol. EEG data acquisition was carried out through 1 channel (using three electrodes) located at points F7 and T3 (and one common reference point in the bone behind the ear) based on the 10-20 system. Because not all participants had short hair, manipulation was carried out by gluing the electrodes at these points using a buff or face covering while driving. The buff was chosen because it has a rubber structure and the area that can cover

the electrode. This data retrieval is also assisted by video media to have a definite time (so that segmentation can be done automatically) and voice recording using the Dolby android application. The application makes it easy for practitioners to denoise recorded voices taken at the same time as the acquisition of EEG data. The experimental setup is shown in FIGURE 2.

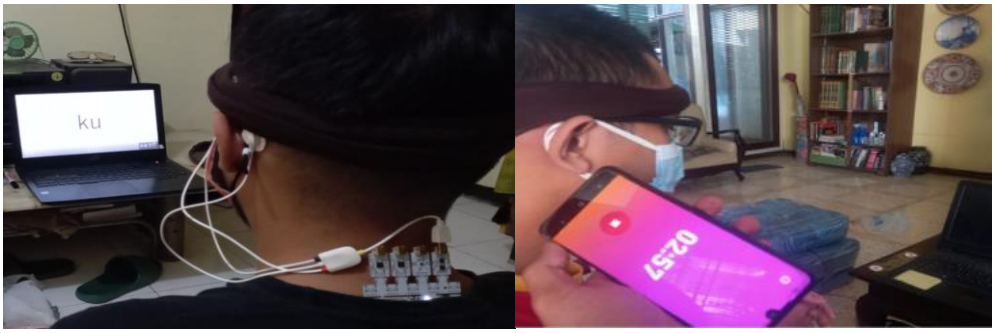


FIGURE 2. Experimental Setup. Collect EEG signal recording through BITalino and PC, as well as voice recording through smartphone.

EEG signal acquisition is done using the same procedure as dataset collection. Furthermore, the data is passed through a filter program in the form of an IIR bandpass filter. This filter has a pass frequency range of 8-30 Hz; inhibition frequencies 1 Hz and 35 Hz; bandpass ripple of 0.4 dB; and the attenuation of the inhibition frequency is 3 dB. This filter is created using the scipy.signal library and the result showed in FIGURE 3. After that, the data that has been obtained is segmented automatically by utilizing patterned recorded sounds. Segmentation is done by cutting one data retrieval into one syllable by one and method (visualization or speech), so that 20 segmented data are produced. The sequence of segmentation methods used is shown in FIGURE 4.

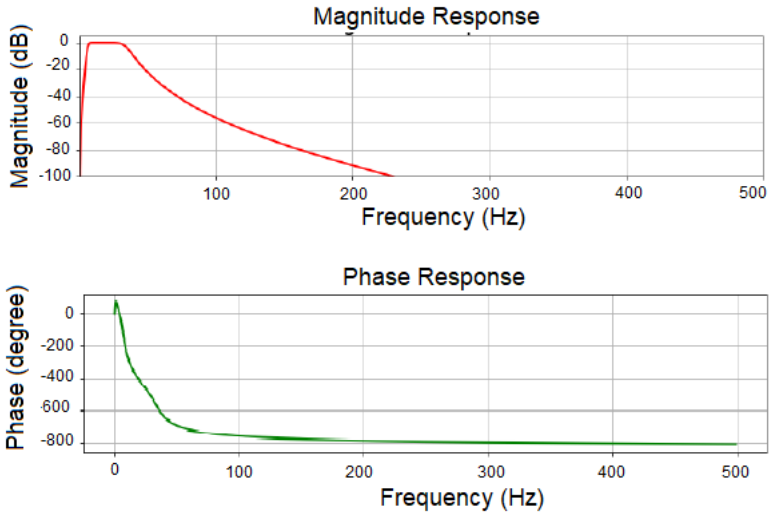


FIGURE 3. The result of IIR Bandpass Filter

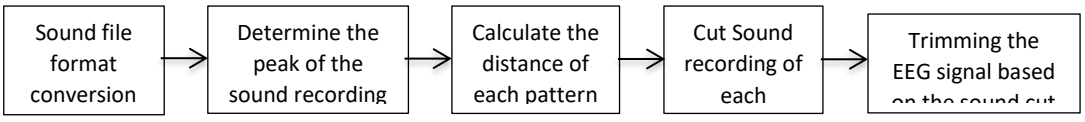


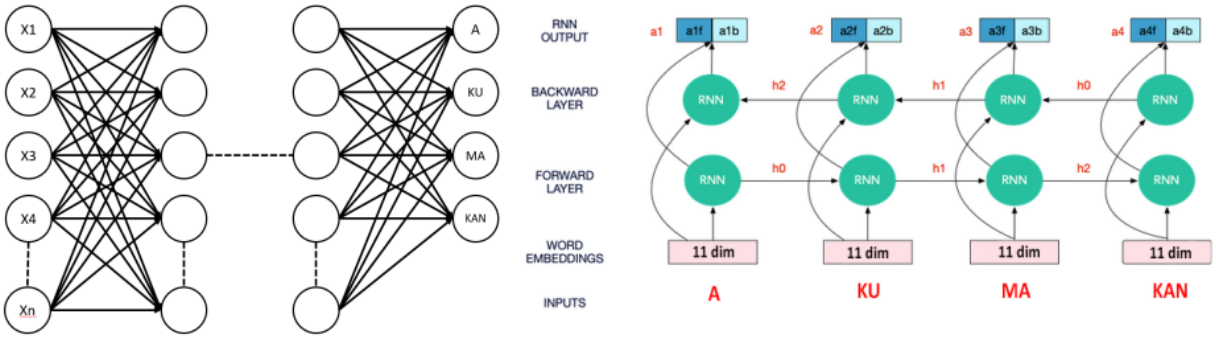
FIGURE 4. The sequence of segmentation methods.

After segmenting, feature extraction is carried out with features in the eeglib library. The features in eeglib used consist of dfa (applying trend fluctuation analysis algorithm), hfd (returning higuchi fractal dimensions), lzc (returning Lempel-Ziv complexity (LZ 76)), pfd (returning petrosian fractal dimensions), hjorth Activity (returning activity Hjorth), hjorth Complexity (returns Hjorth complexity), hjorth Mobility (returns Hjorth mobility), and sampEn (returns sample entropy).

Furthermore, deep learning architecture design is carried out to classify signals and predict the next syllable to be spoken. The output of the deep learning program is text. Those results will then be fed into a text-to-speech program that can convert text into sound. The sound will be issued through a personal computer (and speakers if needed).

There are two deep learning models used, namely artificial neural network (ANN) for classification of EEG signals and recurrent neural network (RNN) type bLSTM to predict what syllables will be uttered by the participants. These two models were built using tensorflow and assisted by sklearn to divide the dataset into 80% training data and 20% test data. In ANN, the layer used is a dense layer. This layer applies backpropagation, so it can correct the resulting error properly. The architecture of the ANN is shown in FIGURE 5(a). The model uses Categorical Cross entropy as the loss function and adam as the optimizer. learning rate and epoch which are considered effective are 0.01 and 100, respectively.

The architecture of the RNN used is not much different from the ANN. The loss function and optimizer used are the same as the ANN. The difference is in the input in the form of text, the number of layers is only three, and there is a bLSTM that is able to consider the results before and after it, so that it can provide more complex considerations. The RNN architecture is shown in FIGURE 5(b).



(a) (b)

FIGURE 5. Established classification model. **(a)** ANN architecture; **(b)** bLSTM architecture.

In this system, 2 models are used in stages, namely ANN and RNN type bLSTM. ANN is used to classify or determine the syllables that the user visualizes. The output of ANN is one syllable. The output is entered as input to the RNN. RNN will predict the next syllable of ANN. The result of RNN is one syllable, so it produces one word (consisting of two syllables). That way, this method has the potential to be developed for assembling sentences as speech synthesis works.

The performance of the best model that we have succeeded in getting an accuracy value of 26.04% with 20.83% validation at epoch 100 and 1000 for ANN by using features in eeglib. This is because there is no overfitting of these features. In addition, 81.25% accuracy was obtained at epochs of 100 and 1000 for the built RNN.

CONCLUSION

Based on the system that has been developed, it can be shown that EEG can be used to form speech synthesis. This is done using acquisition points F7 and T3; IIR filter Butterworth Bandpass pad 8-30 Hz; feature-by-feature extraction on eeglib; 2 stages of deep learning model ANN and RNN type bLSTM; and text-to-speech programs. The accuracy and accuracy of testing on ANN are 26.04% and 20.83%, while the accuracy of RNN is 81.25%. It is recommended to those who want to develop this system to create datasets in the same place and in the same circumstances. In addition, it is also a good idea to collect data at several acquisition points, so that the optimal configuration can be compared. Further research on extraction features and machine learning model comparison are also required to develop optimal accuracy for this purpose.

ACKNOWLEDGMENTS

This study was supported by the Faculty of Science and Technology, Universitas Airlangga (Research grant no. 2417/UN3.1.8/LT/2019). The authors wish to thank the Institute of Research and Innovation (LPI) Universitas Airlangga, faculty of science and Technology Universitas Airlangga and all members who participated in this study.

REFERENCE

1. Al Ansori, A. N. (2020) 'Jumlah Penyandang Disabilitas di Indonesia Menurut Kementerian Sosial', Liputan6.com, 10 September. Available at: https://www.liputan6.com/disabilitas/read/4351496/jumlah-penyandang disabilitas-di-indonesia-menurut-kementerian-sosial#:~:text=Berdasarkan data Susenas pada 2018,atau 30%2C38 juta jiwa.
2. Angrick, M. et al. (2019) 'Speech synthesis from ECoG using densely connected 3D convolutional neural networks', Journal of neural engineering. 2019/03/04, 16(3), p. 36019. doi: 10.1088/1741-2552/ab0c59.
3. Anumanchipalli, G.K., Chartier, J. & Chang, E.F. Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498 (2019). https://doi.org/10.1038/s41586-019-1119-1.
4. Frazer, T.M., 1985. Stress dan Kepuasan Kerja. Terjemahan dari Suwanto. Jakarta: Gramedia.
5. Pasek Suyadnya, I Wayan et al. Alat Bantu Komunikasi Terintegrasi bagi Penyandang Tuna Wicara Berbasis Sensor Gerak dan OpenWrt. Jurnal SPEKTRUM, [S.l.], v. 5, n. 2, p. 176-182, dec. 2018. ISSN 2684-9186.