Non-Invasive Voice Disorder Detection

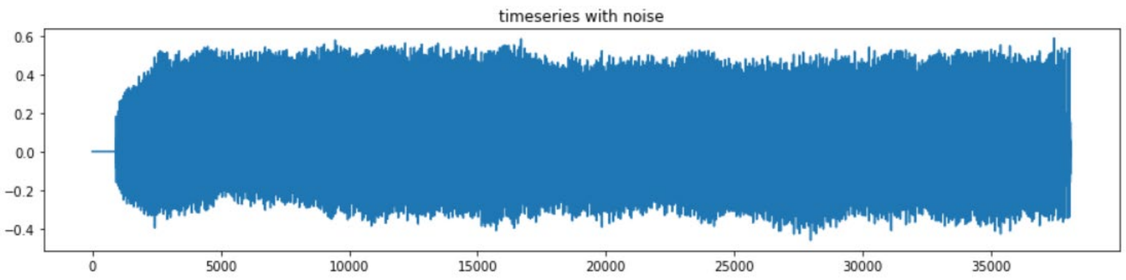
Lim Kee Boon, Gina Tan Ci En, Muhammad Adam
Temasek Polytechnic, Singapore

Abstract

Dysphonia is an alteration of voice production due to a morphological or functional alteration of the pneumo-articulatory apparatus. Moreover, dysphonia affects great number of people, with about one third of adults suffering from this disorder at least once in their lifetime. The conventional method used to diagnose dysphonia is via the use of laryngoscopy. However, the laryngoscopy examination is an invasive procedure performed by an experienced laryngologist. Furthermore, the equipment required is costly and not commonly found in primary care units. The aim of this study is to provide an alternative approach, which is efficient and non-invasive. The project proposed to develop a computer aided diagnosis (CAD) system to classify pathological and healthy voices from voice signal using Multi-layer Perceptron (MLP) and Convolutional neural network (CNN). The 58 healthy and 150 pathological voice signals initially undergone noise removal and downsampled from 8000 Hz to 1000 Hz. This study experimented on both raw and preprocessed voice signals to separately train and test the baseline MLP and 1D CNN models of various layers. Further, the models employed model improvement methods, such as the callbacks, dropout and weight regularization, together with K-fold cross validation and confusion matrix. Overall, 1D CNN model of one layer with callbacks of patience=50 and weight regularization yielded the best validation accuracy of 82.76% and training accuracy of 68.97% at epoch 121.

Materials

CPU: Intel(R) Core(TM) i7-8550u
RAM: 8GB
GPU: GTX 1050

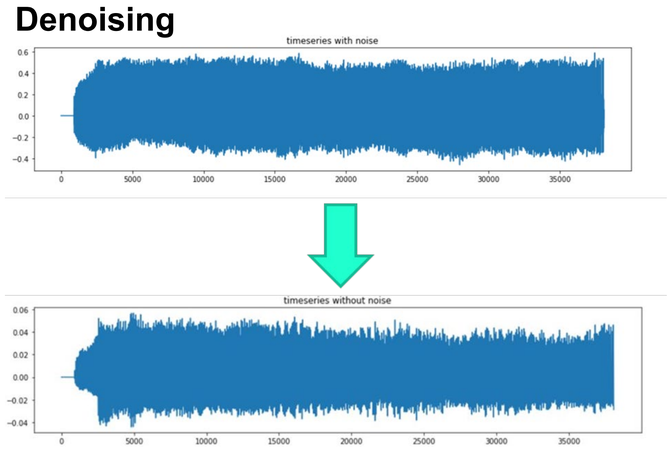


- VOICED (VOice ICar fEDerico II) database
- 150 pathological, and 58 healthy voices.
- Sampled at 8000 Hz

Methods

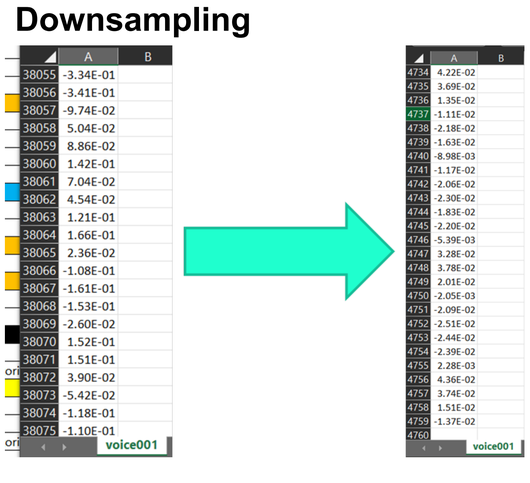
1. Data Pre-processing:

Denoising



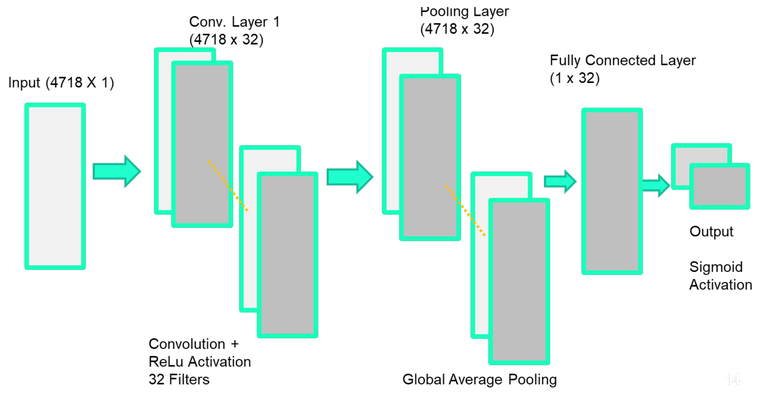
- **Removing the background noise in order to improve the speech signal.**

Downsampling



- **Downsampling from 8000hz to 1000 Hz.**
- **38000 data points to 4700 data points**

2. Neural Network Model:



1D Convolutional Neural Network (CNN) Architecture

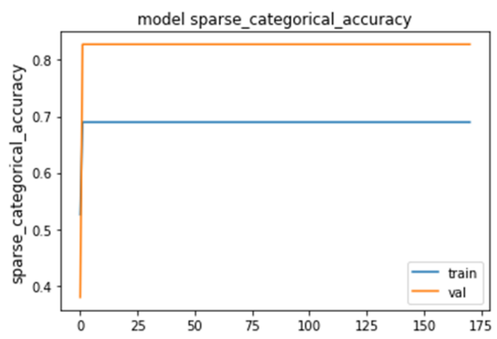
3. Fine tuning:

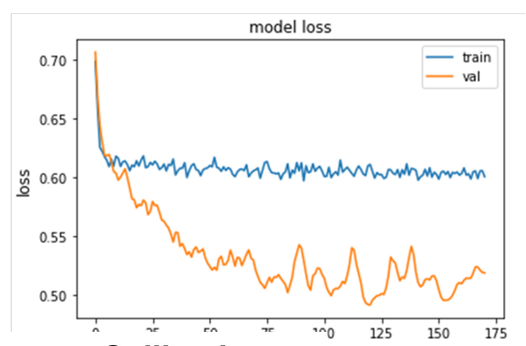
- Dropout
- Weight Decay
- Dropout + Weight Decay

Results & Data Analysis

Best Model Performances

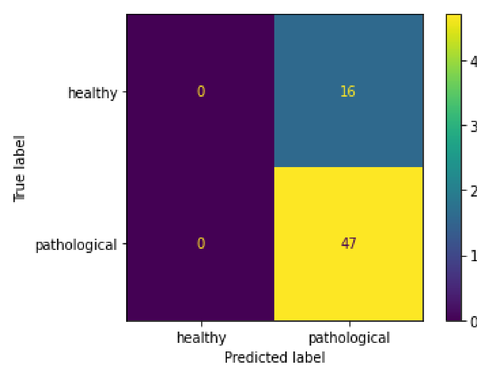
Baseline Model	Optimal Epoch	Training Accuracy (%)	Validation Accuracy (%)	Time Taken for 1 Epoch	Total Time Taken
1	121	68.97	82.76	1s	171s
2	4	75.86	55.17	2s	108s
3	36	68.1	72.41	4s	344s
4	35	81.03	62.07	9s	765s
5	19	80.17	65.52	20s	1380s

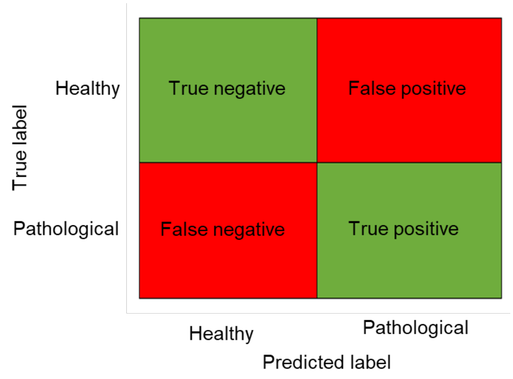




- **Denoised CNN Block 1 + Weight Decay + Callbacks**
- **Accuracy: 82.76%**
- **Remarks: Optimal**

Confusion Matrix





Discussion/Evaluation

In theory, MLP models should have performed better than the CNN models as MLP was more suited for predicting data that is in a tabular format like the csv file we used for our dataset. Yet, the best result was from a CNN model.

The best CNN model of one layer with callbacks of patience=50 and weight regularization yielded the best validation accuracy of 82.76% at epoch 121, which was higher than the best MLP model of 5 layers with a dropout layer of 0.5 which yielded the best validation accuracy of 74.60% at epoch 1. The CNN model performed better than the MLP model most likely due to the limitations of MLP. Since each perceptron in an MLP model is connected to every other perceptron, the number of parameters can grow very high causing redundancy.

If the training accuracy was consistently higher than the validation accuracy, the models would be considered overfitted.

The validation accuracy of all the MLP models with callbacks of patience = 3, both denoised and noise would attain 74.60% at epoch 1. Although the results of the validation accuracies were higher than training accuracies, which may indicate optimal fitting. The accuracy plots however, showed that all the MLP models were overfitted except for denoised MLP model of 5 layers with a dropout layer of 0.5 as well denoised MLP model of 5 layers with a dropout layer of 0.5 and weight regularization, both showing validation accuracies to be consistently higher than training accuracies.
CNN models which had validation accuracy lower than the training accuracy resulted in overfitting. There were 17 different CNN models which had validation accuracy higher than the training accuracy which showed promise in showing optimal fitting. However, 16 of them overfitted and only denoised CNN model of one layer with weight regularization showed validation accuracy consistently higher than training accuracy.

Regarding misclassification, we found that our model was reliable to a certain extent. Even though the model had misclassified 25% of the validation dataset, it had no false negatives. Having false negatives is not ideal as it may cause a delay in treatment.

Conclusions

- Developed a computer aided diagnosis (CAD) system to classify pathological and healthy voices from voice signal using CNN.
- Overall, 1D CNN model of one layer with callbacks of patience = 50 and weight regularization yielded the best validation accuracy of 82.76%.

References

1. Cesari, U., Pietro, G.D., Marciano, E., Niri, C., Sannino, G., & Verde, L. (2018). A new database of healthy and pathological voices. Comput. Electr. Eng., 68, 310-321.
2. Fang, S., Tsao, Y., Hsiao, M., Chen, J., Lai, Y., Lin, F., & Wang, C. (2018). Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. Journal of voice: official journal of the Voice Foundation.
3. Lee, Ji-Yeoun. "Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database." Applied Sciences 11.15 (2021): 7149.
4. Wu, H., Soraghan, J., Lowit, A., & Di Caterina, G. (2018, July). Convolutional neural networks for pathological voice detection. In 2018 40th annual international conference of the ieee engineering.