

Speech Intelligibility Prediction of Dysarthria Using Deep Convolutional Networks

Junseok Oh, Hosung Park and Ji-Hwan Kim*
Department of Computer Science and Engineering, Sogang University
Seoul, South Korea
[e-mail: ohjs@sogang.ac.kr, hosungpark@sogang.ac.kr, kimjihwan@sogang.ac.kr]
*Corresponding author: Ji-Hwan Kim

Abstract

Speech intelligibility, the degree to which a listener understands speech, is a crucial factor in diagnosing and managing speech impairments. This study proposes an objective and efficient method for its prediction using a Convolutional Neural Network, specifically the ResNext101 model. The method relies on a dataset of Diadochokinetic (DDK) task voices from 319 patients, previously assessed by experts. The model takes the Mel-spectrogram of DDK task speech as input and classifies it into five levels of speech impairment. A balanced categorical cross-entropy was used to mitigate data imbalance during model training. The model achieved a macro-average accuracy of 59.6% and a micro-average accuracy of 72.1%, demonstrating its effectiveness in predicting speech intelligibility. This work paves the way for future research aimed at improving accuracy and generalizability across a broader range of speech disorders.

Keywords: speech intelligibility, Convolutional Neural Network, dysarthria

1. Introduction

Speech intelligibility is vital for diagnosing and managing speech impairments [1]. It is affected by factors such as articulation, speech rate, rhythm, and voice quality. Dysarthria, a motor speech disorder, presents significant challenges in speech intelligibility [2]. Expert evaluations, though essential, can be subjective, time-consuming, and costly. Therefore, an automatic prediction method using a Convolutional Neural Network (CNN) model is proposed to classify speech impairment levels [3].

2. Related Works

CNNs have been widely applied in speech and audio processing due to their ability to learn robust features directly from raw data [4]. CNNs have been used in hybrid NN-HMM models for speech recognition, demonstrating their efficiency in learning hierarchical features.

3. Dataset

The dataset[1] consists of recordings from 319 patients with language disorders performing a Diadochokinetic (DDK) task [5]. A DDK task is a standard speech pathology assessment

[1] This research (paper) used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621,Development of artificial intelligence technology that provides dialog-based multi-modal explainability)

involving the rapid, clear repetition of a series of simple syllables, typically /pa/, /ta/, and /ka/. These tasks are used to evaluate the motor abilities of the speech articulators and can reveal potential movement limitations. The patients' performances were previously assessed by experts on a 1-5 speech impairment scale. In total, 957 files were compiled and divided into training, validation, and testing sets at an 8:1:1 ratio.

4. Intelligibility Prediction Model

The ResNext101 [6] architecture with Mel-spectrogram input was employed to predict speech intelligibility automatically. This model incorporates a 2D convolution in the time-frequency domain, enabling it to capture the intricate patterns present in the Mel-spectrogram data efficiently. The model takes a 224-dimensional Mel-spectrogram of DDK task speech as input and classifies it into one of five speech intelligibility levels [7].

To mitigate data imbalance, weighted categorical cross-entropy was used as the loss function. Weighted cross-entropy is a variant of the cross-entropy loss function, which assigns different weights to different classes. This is particularly useful when dealing with imbalanced datasets.

Additionally, data augmentation techniques were utilized to enhance the robustness of the model further. Techniques such as SpecAugment [8] and speed perturbation were employed to create variations in the dataset, thus aiding in better generalization of the model by exposing it to more diverse scenarios within the data.

This combined approach of using a weighted loss function and data augmentation helps effectively deal with data imbalances and enhance the model's performance.

5. Experiment

The model was trained and evaluated using the collected dataset, achieving a macro-average accuracy of 59.6% and a micro-average accuracy of 72.1%. These results suggest the model's effectiveness at predicting speech intelligibility based on DDK task voices.

5. Conclusion

In this paper, we have proposed a CNN-based model for the prediction of speech intelligibility. The proposed model utilizes the ResNext101 architecture, proving itself an efficient, objective tool for automatically predicting speech intelligibility in patients with language disorders. In future work, we will focus on enhancing the accuracy of this model and broadening its generalizability.

References

[1] R. D. Kent *et al.*, "Acoustic studies of dysarthric speech: methods, progress, and potential," *Journal of Communication Disorders*, vol. 32, no. 3, pp. 141-186, 1999.

[2] J. R. Duffy, "Motor Speech Disorders: Substrates, Differential Diagnosis, and Management," Elsevier Mosby, 2005.

[3] S. G. Fletcher, "Time-by-count measurement of diadochokinetic syllable rate," *Journal of Speech and Hearing Research*, vol. 15, no. 4, pp. 763-770, 1972.

[4] O. Abdel-Hamid *et al.*, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *in Proc. of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277-4280, 2012.

[5] S. G. Fletcher, "Time-by-count measurement of diadochokinetic syllable rate," *Journal of Speech and Hearing Research,* vol. 15, no. 4, pp. 763-770, 1972.

[6] S. Xie *et al.*, "Aggregated Residual Transformations for Deep Neural Networks," *in Proc. of the IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 1492-1500, 2017.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[8] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *in Proc. of the Interspeech 2019*, pp. 2613-2617, 0Graz, Austria, Sep. 2019.