Pathological Speech Classification Using a Convolutional Neural Network

Nam H. Trinh, Darragh O'Brien

ADAPT Centre, School of Computing, Dublin City University, Ireland

Abstract

Convolutional Neural Networks (CNNs) have enabled significant improvements across a number of applications in computer vision such as object detection, face recognition and image classification. An audio signal can be visually represented as a spectrogram that captures the time-varying frequency content of the signal. This paper describes how a CNN can be applied to the spectrogram of an audio signal to distinguish pathological from healthy speech. We propose a CNN structure and implement it using Keras to test the approach. A classification accuracy of over 95% is obtained in experiments on two public pathological speech datasets.

Keywords: Pathological Speech, Audio Classification, Spectrogram, Convolutional Neural Network.

1 Introduction

In recent years, deep learning applications in healthcare have attracted considerable research attention. Such applications enable early clinical diagnoses from X-ray images or patients' speech data. One application is the use of deep neural networks to classify speech data as pathological or healthy. Such pathologies include neurological diseases such as Parkinson's or Alzheimer's disease. Several methods have been proposed to distinguish between Parkinsonian and healthy speech such as [Asgari and Shafran, 2010],[Sakar et al., 2013], [Tsanas et al., 2012] and [Viswanathan et al., 2018]. CNNs have achieved considerable success in the field of computer vision and a CNN approach to pathological speech detection has recently been proposed by [Alhussein and Muhammad, 2018].

In addition to reviewing relevant work in the area, this paper further explores the recently proposed CNN-based method for classifying pathological speech by, firstly, testing an alternative CNN structure to that of [Alhussein and Muhammad, 2018] and, secondly, by validating the CNN approach against two independent speech datasets. The paper is organised as follows. In Section 2 an overview of related work in the area of pathological speech classification is presented that includes a description of associated features, methods and datasets. Section 3 describes our methodology and our proposed CNN model. Results are presented in Section 4 and Section 5 concludes.

2 Related Work

Selected related work is summarised in Table 1. A typical Pathological Speech Classification (PSC) model includes two main components: a feature extractor for speech signal processing that computes salient features and a classifier (illustrated in Figure 1). Some early work used a Support Vector Machine (SVM) as a classifier with Mel-Frequency Cepstral Coefficients (MFCCs) as input features. [Poorjam et al., 2018] proposed this approach and achieved an accuracy of 88%.

With the emergence of deep learning algorithms, PSC models based on neural networks have also been proposed. [Moon and Kim, 2018], [Smitha et al., 2018], [Fang et al., 2018] and [Shia and Jayasree, 2017] used a Multilayer Perceptron (MLP) in their work. In their work, MFCCs served as input vectors to the MLP. MLP drawbacks however, include overfitting and a long training time due to the large number of model parameters.

To address MLP issues, CNN-based models were proposed. Using a CNN-based approach[Alhussein and Muhammad, 2018] achieved a state-of-the-art result (97.5% accuracy).

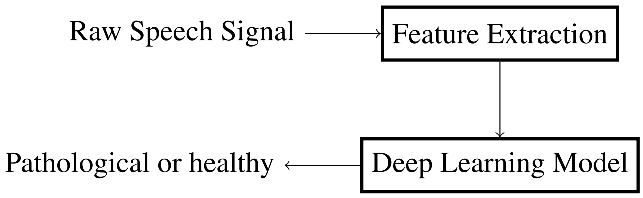

Figure 1: PSC Model

Table 1: Related work in pathological speech classification: features, classifiers and reported accuracy

Reference paper	Dataset	Features	Classifier	Accuracy
[Poorjam et al., 2018]	Data collected by the authors in collaboration with Sage Bionetworks and PatientsLikeMe	MFCCs	SVM	88.0%
[Moon and Kim, 2018]	SVD	Jitter, shimmer and MFCCs	MLP	87.4%
[Smitha et al., 2018]	Data supplied by Speech Language Pathologist from the Nitte Institute of Speech and Hearing Mangaluru	MFCCs	MLP	95.0%
[Fang et al., 2018]	Data from FEMH challenge with 60 normal voice samples and 402 pathological voice samples	MFCCs and MFCCs + delta features	MLP	94.5%
[Shia and Jayasree, 2017]	SVD	Wavelet Subband Energy Coefficients	MLP	93.3%
[Harar et al., 2017]	SVD	No extracted features (raw signal)	CNN + RNN with LSTM	68.1%
[Wu et al., 2018]	SVD	Spectrogram	CNN	71.0%
[Alhussein and Muhammad, 2018]	SVD	Spectrogram (after framing and applying STFT)	CNN	97.5%

3 Methodology

Given in Figure 2 are example spectrograms extracted from pathological and healthy speech samples. Distortion across the pathological speech sample is observed. By contrast, the frequency content of spectrograms from healthy speech samples is more stable. The goal is to use a CNN to detect the distortions and instabilities in the spectrogram indicative of pathological speech.

In our work, spectrograms are extracted using Librosa [McFee et al., 2015]. The speech signals are first windowed (with a window length of 25ms) and the Short-time Fourier Transform (STFT) is subsequently applied to extract the frequency components of the audio signal. The resulting image is fed to a CNN for classification.

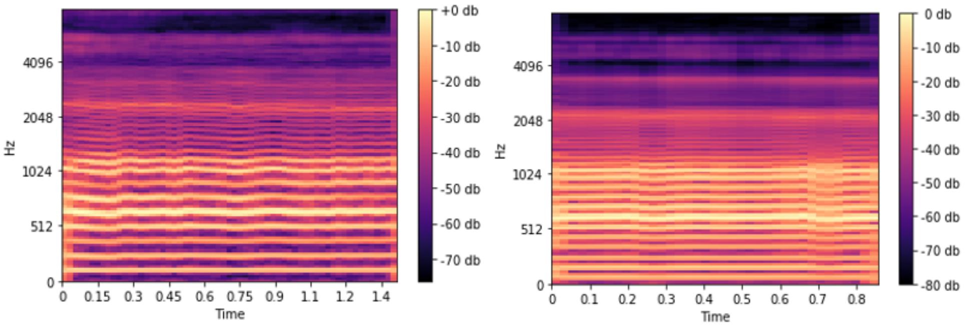

Figure 2: Spectrograms of a pathological speech sample (left) and of a healthy speech sample (right)

The architecture of the CNN model used in this work is summarised in Figure 3. The input layer has shape 28x28x3. The model contains three convolutional layers, one max-pooling layer, two fully-connected layers and one output layer. The first convolutional layer has 16 filters of size 3x3, with a same padding and a stride of one followed by a batch normalization layer [Ioffe and Szegedy, 2015]. The second convolutional layer has 32 filters of size 3x3, with a same padding and a stride of one followed by a batch normalization layer. The third convolutional layer has 64 filters of size 3x3, with a same padding and a stride of one followed by a batch normalization layer. A max-pooling layer with a size of two and a stride of two follows the convolutional layers and shrinks the size of the data by a factor of two. This layer's output is flattened and fed into two fully-connected layers with 128 and 64 neurons. The final output layer is a single neuron for binary classification with a sigmoid activation function. The total number of parameters is 1,638,113. We use Keras on top of Tensorflow to build the model.

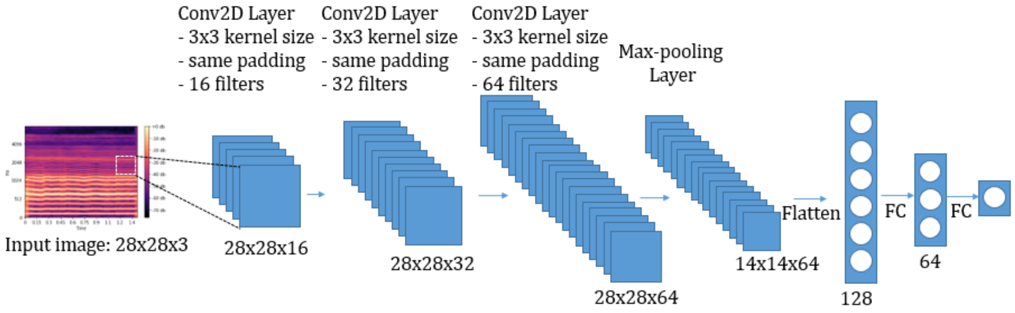

Figure 3: CNN architecture

4 Experiments and Results

The proposed CNN was tested against two datasets, namely, the Saarbrucken Voice Database [Barry and Pützer, 2007] and the Spanish Parkinson's disease dataset [Orozco-Arroyave et al., 2014]. Below we describe both the datasets and the results.

4.1 Datasets

The Saarbrucken Voice Database (SVD): SVD is a collection of speech samples from more than 2000 people. There are three types of recording in the dataset:

- Recordings of sustained vowel sounds (/a/,/u/ and /i/) at normal, high and low pitch,
- recordings of sustained vowel sounds (/a/, /u/ and /i/) at rising-falling pitch,
- recordings of a conversational sentence in German.

In our work, we use a subset of SVD composed of 50 pathological speech samples and 53 healthy speech samples of the sustained /a/ vowel. Multiple samples are extracted from each file.

The Spanish Parkinson's Disease Dataset (SPDD): SPDD includes speech samples from 50 Parkinson's disease patients and 50 healthy controls. Several types of speech are included:

- Recordings of sustained vowels in Spanish,
- recordings of some specific words and phonemes,
- recordings of three sets of different words,
- recordings of conversational speech.

As with SVD, we use the sustained vowel /a/ recordings to test our model and to compare its performance across two independent datasets.

4.2 Results

The CNN was trained using the Adam Optimizer [Kingma and Ba, 2014], the minibatch size was 32, the number of epochs was 30. The CNN was trained on 80% of each dataset and tested against the remaining 20% of that dataset.

The performance of the model on these datasets is summarised in Table 2. Results show that the model achieves 99% test accuracy on SVD, which is competitive with that reported by [Alhussein and Muhammad, 2018] (see Table 1). With SPDD, we achieve a 98.84% training accuracy and 96.70% test accuracy. The gap between training accuracy and test accuracy indicates some overfitting when training the model with this dataset.

Table 2: Achieved results with different datasets

Dataset	Training Accuracy	Test accuracy
SVD	99.81%	99.00%
SPDD	98.84%	96.70%

5 Conclusion

This paper describes our further study of the recently proposed CNN-based approach to pathological speech classification in which spectrograms extracted from raw audio signals are fed into a CNN as input images. The CNN model was built with three convolutional layers, one max-pooling layer and two fully-connected layers using Keras on top of Tensorflow. The classification accuracy on the SVD was 99% which is comparable with other state-of-the-art results in pathological speech classification using the same dataset. The model was further validated against another, independent dataset, SPDD, where a test accuracy of 96.70% was achieved. Our results confirm that the image-based approach using spectrograms as CNN inputs can be used to classify speech signals with high accuracy.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Nam H. Trinh is funded by Science Foundation Ireland under grant No. 17/RC/PHD/3488.

References

[Alhussein and Muhammad, 2018] Alhussein, M. and Muhammad, G. (2018). Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*, 6:41034–41041.

[Asgari and Shafran, 2010] Asgari, M. and Shafran, I. (2010). Extracting cues from speech for predicting severity of parkinson's disease. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 462–467. IEEE.

[Barry and Pützer, 2007] Barry, W. and Pützer, M. (2007). Saarbrucken voice database. *Institute of Phonetics, Universität des Saarlandes, http://www. stimmdatenbank. coli. uni-saarland. de*.

[Fang et al., 2018] Fang, S.-H., Tsao, Y., Hsiao, M.-J., Chen, J.-Y., Lai, Y.-H., Lin, F.-C., and Wang, C.-T. (2018). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*.

[Harar et al., 2017] Harar, P., Alonso-Hernandezy, J. B., Mekyska, J., Galaz, Z., Burget, R., and Smekal, Z. (2017). Voice pathology detection using deep learning: a preliminary study. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–4.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[McFee et al., 2015] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

[Moon and Kim, 2018] Moon, J. and Kim, S. (2018). An approach on a combination of higher-order statistics and higher-order differential energy operator for detecting pathological voice with machine learning. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 46–51.

[Orozco-Arroyave et al., 2014] Orozco-Arroyave, J. R., Arias-Londoño, J. D., Bonilla, J. F. V., Gonzalez-Rátiva, M. C., and Nöth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *In Proc. Of the International Confer- ence on Language Resources and Evaluation (lrec)*, pages 342–347, Reykjavik, Iceland.

[Poorjam et al., 2018] Poorjam, A. H., Little, M. A., Jensen, J. R., and Christensen, M. G. (2018). A parametric approach for classification of distortions in pathological voices. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

[Sakar et al., 2013] Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., and Kursun, O. (2013). Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834.

[Shia and Jayasree, 2017] Shia, S. E. and Jayasree, T. (2017). Detection of pathological voices using discrete wavelet transform and artificial neural networks. In *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pages 1–6.

[Smitha et al., 2018] Smitha, Shetty, S., Hegde, S., and Dodderi, T. (2018). Classification of healthy and pathological voices using mfcc and ann. In *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–5.

[Tsanas et al., 2012] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE transactions on biomedical engineering*, 59(5):1264–1271.

[Viswanathan et al., 2018] Viswanathan, R., Khojasteh, P., Aliahmad, B., Arjunan, S., Ragnav, S., Kempster, P., Wong, K., Nagao, J., and Kumar, D. (2018). Efficiency of voice features based on consonant for detection of parkinson's disease. In *2018 IEEE Life Sciences Conference (LSC)*, pages 49–52. IEEE.

[Wu et al., 2018] Wu, H., Soraghan, J., Lowit, A., and Di Caterina, G. (2018). A deep learning method for pathological voice detection using convolutional deep belief networks. In *Interspeech 2018*.