Design and Validation of a New Diagnostic Tool for the Differentiation of Pathological Voices in Parkinsonian Patients

Eleana E.I. Almaloglou[1], Stella Geronikolou[2[2]], George Chroussos[2[3]], and Constantine Kotropoulos[1[0000-0001-9939-7930]]

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{elenalma, costas}@csd.auth.gr
[2] Clinical, Experimental Surgery and Translational Research Centre, Biomedical Research Foundation Academy of Athens, Athens, Greece
sgeronik@bioacademy.gr

Pathological speech, in its many forms, is a symptom of numerous serious diseases affecting millions of people worldwide, including more than 10 million Parkinson patients. Here, a powerful method is proposed for detecting pathological speech, using a two-dimensional (2D) convolutional neural network (CNN). Spectrograms are extracted from voice recordings of healthy and Parkinson diagnosed patients, which are fed into the CNN architecture. The voice samples comprise a subset of the benchmark mobile Parkinson Disease (mPower) study. The proposed model achieves 98% accuracy in Parkinson detection (i.e., a two-class problem). Moreover, an average accuracy exceeding 94% is measured in binary tests (i.e., pathological versus healthy) employing 6 voice pathologies conducted on the Saarbrucken voice database. These pathologies are dysphonia, functional dysphonia, hyper functional dysphonia, spasmodic dysphonia, vocal fold polyp, and dysody.

Keywords: Pathological Speech, Deep Learning, Audio Classification, Spectrogram, Convolutional Neural Network, mPower study, Saarbrucken voice database.

1 Introduction

Voice disorders afflict millions of people and can cause from mild discomfort to serious pain and loss of communication. A large number of these disorders are due to irregularities in the vocal folds. Vocal folds are membranes in the larynx, which control air circulation and vibrate in order to produce voiced sounds. The vibration frequency range varies from 60 Hz to 300 Hz. The uniqueness of each person's voice is determined by the fundamental frequency, which is the main contributor to the perception of pitch in speech [1].

Voice change is one of the secondary motor symptoms of Parkinson disease (PD) [2]. As PD progresses, movement of various parts of the body is affected and muscles get harder to control. Similar effects are noticed in the vocal folds of PD patients. Many PD patients suffer from dysarthria, a motor speech disorder affecting patient's articulation, phonation, prosody, and respiration. As vocal folds weaken, the voice may get hoarser, breathy, and speech may be slurred and interrupted by long pauses [3]. Other pathologies, such as dysphonia and vocal cysts affect the physiology of the vocal tract in similar manner with PD [4]. With traditional diagnostic procedures being costly and time consuming, a simple method is greatly needed to detect accurately emerging pathologies. Changes in the vocal folds related to spectrum can be pinpointed using spectrogram analysis, which gives the opportunity to offer a quick and easy way to detect possible pathological phenomena and inform patients.

Advances in signal analysis and speech processing have introduced novel solutions to the diagnostic process. For example, the analysis of voice recordings is an effective, non-invasive diagnostic process. Knowledge gained by state-of-the-art techniques in speech recognition has raked gains in voice pathology detection. For example, linear prediction coefficients extracted from utterances were used as features and the receiver operating characteristic (ROC) curve of the linear classifier for vocal fold paralysis and vocal fold edema detection was derived in [5]. Such a linear classifier stemmed from the Bayes classifier, when Gaussian class conditional probability density functions with equal covariance matrices were assumed. The optimal operating point of the linear classifier was specified with and without reject option. The reject option was shown to yield statistically significant improvements in the accuracy of detecting the voice pathologies under study [5]. Other frequently used features are the Mel-Frequency cepstral coefficients (MFCCs). MFCCs can simulate the human hearing mechanism but are not easily interpretable in relation to laryngeal physiology [6]. In [7], feature extraction methods from the Audio-Visual Emotion Recognition Challenge [8] were used with a Gradient Boosted Decision Tree classifier for Parkinson's diagnosis (i.e., a two-class problem) on the mobile Parkinson Disease (mPower) database [9]. An accuracy of 86% was reported. A neurocomputational framework was presented to model the speech production process [10]. The model generated biomarkers of disease by modeling vocal source control with two muscle parameters and their coordination. The derived features were applied to both the Audio-Visual Emotion Recognition Challenge and the mPower databases. A Fisher vector image representation method was tested on the PC-GITA Spanish language dataset to encode the gradients of the log-likelihood of features under the Gaussian Mixture Models. That method combined with a Support Vector Machine (SVM) classifier resulted in 84% accuracy on Parkinson detection of pathological speech versus (vs.) healthy [11].

The rapid progress of Deep Neural Networks brought forth novel approaches to speech analysis. Among the deep learning architectures, the ones with the most discriminative potential are the Convolutional Neural Networks (CNNs) [12]. CNNs are fast and very efficient in image classification tasks and require little preprocessing, which results in a reduction of the human error factor. Advantages like these, make them ideal for medical image analysis, like breast cancer [13] and diabetic retinopathy detection [14].

In this paper, a CNN is used to detect pathological from healthy voices. As input to the CNN, mel-spectrograms are extracted from voice samples recorded by a smartphone. A mel-spectrogram captures the resonating frequencies of vocal tract and their variation in the temporal domain. It unlocks the vocal tract information of a pathological physiology [6]. The proposed model succeeds to detect Parkinson patients from voice recorded in ambient conditions. Moreover, the proposed model, trained for detecting PD by processing mPower recordings, can also detect 6 voice pathologies (i.e., healthy vs. pathological) in experiments conducted on recordings from Saarbrueken Voice Database (SVD). These pathologies are dysphonia, functional dysphonia, hyper functional dysphonia, spasmodic dysphonia, vocal fold polyp, and dysody.

The rest of the paper is organized as follows: Section 2 describes the datasets used. Section 3 analyzes data preprocessing steps and discusses the proposed model. Section 4 describes the experiments on different datasets and discloses experimental findings. Conclusions are drawn in Section 5.

2 Dataset Description

2.1 mPower

Data are extracted from the mPower database [9]. The database is the outcome of a study consisting of four activities ('memory', 'tapping', 'voice' and 'walking') that the participants could complete three times a day. A mobile app available in the Apple store was developed to collect recordings from users, which made this study accessible to 5,826 unique participants, both male and female, from the comfort of their own home. By simply downloading the app and filling out a demographic study, each individual could complete any activity, resulting in an astounding volume of data.

This paper is focusing on "voice recordings". Using their personal smartphone, the participants recorded themselves trying to sustain the sound 'aah' for as long as possible. A recording of the background noise was required prior to the activity to ensure the best data quality. Overall, 65,022 voice recordings, of 10 seconds (s) maximum duration each, were collected. 61,482 were finally used in the experiments here, after removing defective files. The size of the dataset makes it an excellent candidate for training and testing deep learning methods.

2.2 Saarbrueken Voice Database

SVD [15] is a collection of voice recordings from healthy and pathological participants with various forms of voice pathologies. Each entry of the database consists of the following recordings:

• Recordings of sustained vowel sounds [i, a, u] produced at normal, high, and low pitch,

• Recordings of sustained vowel sounds [i, a, u] with rising-falling pitch,

• Recordings of the sentence "Guten morgen, wie geht es lhmen?" (Good morning! How are you?).

All audio files are about 1.5 s long. Here, 6 SVD subsets are used to test the trained PD voice pathology detection model, using all intonations of the sustained vowel 'a' recordings. In order to report comprehensive results, a minimum of 100 pathological recordings were considered as a qualifying criterion for the choice of pathologies, as shown in Table 1.

Table 1. SVD Pathologies.

PATHOLOGY	NUMBER OF SAMPLES	
	Healthy	***Pathological***
Dysphonia	300	303
Functional Dysphonia	300	336
Hyper Functional Dysphonia	600	639
Spasmodic Dysphonia	600	612
Vocal Fold Polyp	100	135
Dysody	100	168

3 Methods

3.1 Preprocessing of input data

First the voice recordings are converted into wav format. The recordings are trimmed using Librosa [16] to maintain 1.5 s long audio files from the middle of each recording to avoid end effects (e.g., background noise). Audio recordings of duration smaller than 1.5 s are removed from the dataset along with recordings with long silence intervals. Short-time Fourier transform (STFT) is applied to 92 ms long windows in order to represent the signal in the time-frequency domain. Next, the Mel-spectrogram is extracted. The Mel-spectrogram was chosen due to the ability of Mel scale to match the way the human auditory system works. Most of the audio processing was done using Librosa [16]. Finally, the Mel-spectrograms are resized to 64x64 RGB images. The difference between a healthy and a pathological Mel-spectrogram is demonstrated in Figure 1. The effects of jitter (change of pitch) and shimmer (change of amplitude) [17] in the time-frequency distribution of the power for a pathological voice can easily be observed.

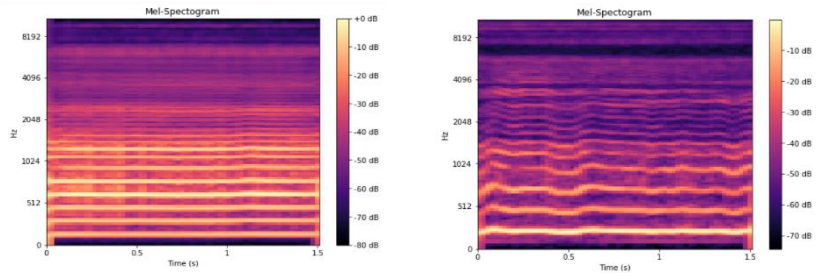

Fig. 1. Healthy (left) and pathological (right) voice mel-spectrograms.

3.2 Proposed model

The proposed deep neural network consists of 7 layers. The first 3 convolutional layers followed by a maxpooling layer comprise the feature learning part of the network. The flattened feature matrix enters 2 fully connected (FC) layers and a final output sigmoid layer. The complete model is shown in Figure 2.

Exponential Linear Unit activation function, same padding, and stride of 1 was used for all convolutional layers. The first and second layers have 16 and 32 filters, with kernel size (5, 5) respectively. The third layer has 64 filters with kernel size (3, 3). The output of the maxpooling layer of pool size (2, 2) is flattened into a vector of size (65,536) and fed into the first FC layer of 128 units with a Rectified Linear Unit activation function. A second FC layer of 64 units is connected with a binary sigmoid neuron, forming the output layer.

The Adam optimizer was chosen with a learning rate of 0.001 and an epsilon value of 1e-7. The initial weights were set randomly with a Glorot uniform initializer and the loss function used was binary cross-entropy. The total number of trainable parameters is 8,429,666. The parameters of the deep neural network were chosen after a multitude of experiments. The framework was developed in python 3.7 using Keras on top of TensorFlow.

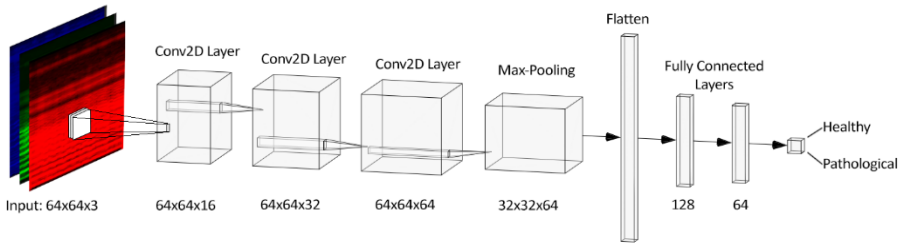

Fig. 2. Model Architecture.

4 Experimental Results

The mPower dataset was split into 3 subsets, namely, a training set (56%), a validation one (14%), and a test set (30%), where the number inside parentheses refers to the proportion of data used. The train and validation sets are both labeled. The model learns on the training data and adjusts the weights using back propagation. Simultaneously, it applies the weights of each epoch on the validation data with simple forward propagation. This process allows to monitor model ability to generalize and not overfit. The validation process does not alter model weights. Early stopping is employed to avoid unnecessary training. The model was trained for 12 epochs and achieved a validation accuracy of 98,5%. The trained model architecture and weights are saved using Keras. The saved model is used to predict the labels of the test set and the unique speaker set for the final evaluation. A 4th subset derived from mPower is the so-called unique speaker set. It was created in order to ensure that the model is not sensitive to speaker-identity. It consists of 1215 audio recordings from 230 unique participants not included in the training, validation, or test set.

Metrics of accuracy (*ACC*), specificity, sensitivity, and precision were calculated based on the classification results. Sensitivity or true positive rate (*TPR*) reveals the ability of the model to detect the truly pathological patients (i.e., patients who are diagnosed with a pathology and classified by the model as pathological). Specificity or true negative rate (*TNR*) reveals the ability of the model to identify the healthy subjects (i.e., patients who are healthy and are classified as such by the model). Their formal definitions are:

$$ACC = (TP+TN)/ (TP + TN + FP + FN) \quad (1)$$

$$TPR = TP / (TP + FN) \quad (2)$$

$$TNR = TN / (FP + TN) \quad (3)$$

where *TP* is the number of true positive samples (i.e., actually pathological classified as pathological), *TN* is the number of true negative samples (i.e., actually healthy classified as healthy), *FN* is the number of false negative samples (i.e., actually pathological classified as healthy), and *FP* is the number of false positive samples (i.e., actually healthy classified as pathological). Empirical measurements are summarized in Table 2.

Table 2. Model Accuracy on mPower.

Training Accuracy	Validation Accuracy	Test Accuracy
98.7%	98.5%	98%

In Tables 3 and 4, the confusion matrix and the figures of merit on the test set are listed.

Table 3. Confusion matrix on the test set.

	True Pathological	**True Healthy**
Predicted Pathological	11,381	170
Predicted Healthy	171	6,723

Table 4. Evaluation on test set.

No. of samples	**Accuracy**	**Sensitivity**	**Specificity**
18,445	98%	98.5%	97.5%

The trained model was further tested on the unique speaker set. It was crucial to attest model's lack of bias, since mPower comprises multiple recordings of each subject. The confusion matrix and model performance evaluation on the unique speaker set are shown in Tables 5 and 6. The 98% accuracy on the unique speaker set confirms that the model is well trained to detect pathological voices and not to perform speaker verification. In both the test set and the unique speaker set, the model does not exhibit any bias in favor of any class.

Table 5. Confusion matrix on the unique speaker set.

	True Pathological	**True Healthy**
Predicted Pathological	599	12
Predicted Healthy	13	591

Table 6. Evaluation on unique speaker set.

No. of samples	**Accuracy**	**Sensitivity**	**Specificity**
1,215	98%	97.8%	98%

To assess the diagnostic ability and confidence of the CNN classification, the ROC curve was calculated for each set. The area under the curve (AUC) was calculated as an additional figure of merit. All tests resulted in an AUC exceeding 0.989. Due to lack

of space, the ROC curve of PD detection on the unique speaker set is only exhibited in Figure 3.

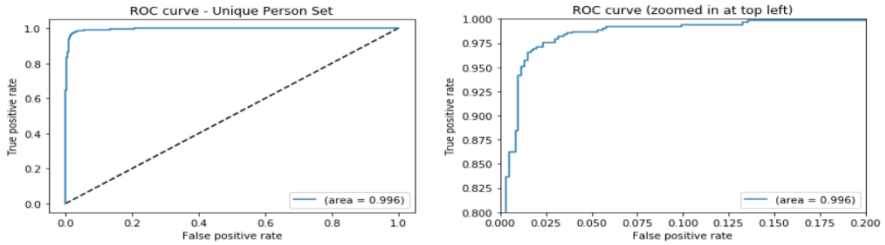

Fig. 3. ROC curve for the unique speaker set.

Afterward the trained voice pathology detection model was tested separately on 6 different pathologies from SVD whose clinical image is similar to that of PD. Two-class classification problems were conducted (i.e., pathological vs. healthy speech). The results are listed in Table 7. The detection accuracy of most pathologies is remarkably close to that on the mPower database, on which the model was trained, with dysody and hyper functional dysphonia achieving the highest accuracy.

Table 7. Accuracy of detection on SVD pathologies.

	Accuracy	**Specificity**	**Sensitivity**
Dysphonia	96.2	96.4	96.1
Functional Dysphonia	96.3	96.1	96.4
Hyper Functional Dysphonia	97.1	96.4	97.6
Spasmodic Dysphonia	94	94	93.5
Vocal Fold Polyp	96.2	93.3	98.4
Dysody	97.4	95.1	98.8

Here, each SVD pathology is tested separately. This allows for a better insight on the performance of the model and its use for pathological voice detection across databases. The accuracy of detecting various pathologies varies up to 3.4%. Special attention is advised. For example, it is observed that sensitivity is elevated in the cases of vocal fold polyp and dysody, which results in greater confidence in pathological diagnosis. If said pathologies where tested as one class, those nuances would be lost.

Although there have been many studies on voice pathology detection using the SVD database, the current results cannot be directly compared to the results appearing in the literature, since it is a unique cross-database experiment where the model is trained on an entirely different dataset with specific pathologies. Table 8 summarizes accuracies on pathological voice detection reported on SVD for related works to setup the landscape.

Table 8. Related work on SVD.

	Method	Accuracy %
Harar et al.,2017 [18]	Raw signal + CNN+RNN	68.1
Wu et al., 2018 [19]	Spectrogram + CNN	71.0
Alhussein and Muhammad, 2018 [20]	Spectrogram + CNN	97.5
S. Roy et al., 2019 [21]	MFCC + CNN	75.64

5 Conclusions

In this paper, a model for differentiating pathological voices from healthy ones is proposed based on a CNN, applied to raw speech Mel-spectrograms. The model was trained and tested on the mPower study dataset and further tested on various SVD pathologies. Promising empirical findings are disclosed in a cross-database experiment. This enables voice pathology detection independent of recording devices or language. Future research could include gender and age information in the experimental protocol. EEG signal could also be fed as secondary input on the neural network architecture using fussion techniques.

6 Acknowledgements

Data was contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [9].
The authors are also grateful to Manfred Pützer and William J. Barry, Institut für Phonetik, Universität des Saarlandes for granting access to SVD [15].

References

1. Huang, X., Acero, A., Hon, H-W.: Spoken Language Processing: A guide to Theory, Algorithm, and System Developement. 1st edn. Prentice Hall, New Jersey (2001).
2. Mosley, A. D., Romaine, D. S.: The Encyclopedia of Parkinson's Disease. 1st edn. Facts on File, New York (2004).
3. Factor, S. A., Weiner, W. J.: Parkinson's Disease: Diagnosis and Clinical Management. 2nd edn. Demos, New York (2008).
4. Schwartz, S. R., Cohen, S. M., Dailey, S. H., Rosenfeld, R. M., Deutsch, E. S., Gillespie, M. B., Patel, M. M.: Clinical Practice Guideline: Hoarseness (Dysphonia). Otolaryngology–Head and Neck Surgery (141) (1_suppl), 1-31, https://doi.org/10.1016/j.otohns.2009.06.744 (2009).

5. Kotropoulos, C., Arce, G. R.: Linear discriminant classifier with reject option for the detection of vocal fold paralysis and vocal fold edema. EURASIP Advances in Signal Processing (2009), article ID 203790, https://doi.org/10.1155/2009/203790 (2009).
6. Lee, J. W., Kang, H. G., Choi, J. Y., Son. Y. I.: An investigation of vocal tract characteristics for acoustic discrimination of pathological voices. BioMed Res Int. (2013), 758731 (2013).
7. Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., Ghomi, R. H.: Parkinson's Disease Diagnosis Using Machine Learning and Voice. In: 2018 IEEE Signal Processing in Medicine and Biology Symposium, pp. 1-7. IEEE, Philadelphia, PA (2018).
8. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In: 3rd ACM Int. Workshop Audiovisual Emotion Challenge, pp. 3–10. Association for Computing Machinery, New York, USA (2013).
9. Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., Friend, S. H., Trister, A. D.: The mPower Study, Parkinson Disease Mobile Data Collected Using Researchkit. Scientific Data (3), 160011 (2016).
10. Ciccarelli, G., Quatieri, T. F., Ghosh, S. S.: Neurophysiological Vocal Source Modeling for Biomarkers of Disease. In: 17th Annual Conf. Int. Speech Communication Association INTERSPEECH 2016, pp. 1200-1204, ISCA, San Francisco, USA (2016).
11. López, J. V. E., Orozco-Arroyave, J. R., Gosztolya, G.: Assessing Parkinson's Disease from Speech Using Fisher Vectors. In: 20th Annual Conf. Int. Speech Communication Association INTERSPEECH 2019, pp. 3063-3067, ISCA, Graz, Austria (2019).
12. Fayek, H. M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for Speech Emotion Recognition. Neural networks : the official journal of the International Neural Network Society (92), 60-68 (2017).
13. Dabeer, S., Khan, M. M., Islam, S.: Cancer diagnosis in histopathological image: CNN based approach. Informatics Med. Unlocked (16), 100231 (2019).
14. Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., Zheng, Y.: Convolutional neural networks for diabetic retinopathy. Procedia Computer Science (90), 200–205 (2016).
15. Putzer, M., Barry, W. J.: Saarbrucken Voice Database. Institute of Phonetics, Universität des Saarlandes (2007). http://www.stimmdatenbank.coli.uni-saarland.de
16. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: Librosa: Audio and Music Signal Analysis in Python. In: 14th Python in Science Conference, pp. 18-24. SciPy, Austin, Texas, USA (2015).
17. Teixeira, J. P., Oliveira, C., Lopes, C.: Vocal Acoustic Analysis-Jitter, Shimmer and HNR Parameters. Procedia Technology (9), 1112–1122 (2013).
18. Harar, P., Alonso-Hernandezy, J. B., Mekyska, J., Galaz, Z., Burget, R., Smekal, Z.: Voice pathology detection using deep learning: a preliminary study. In: 2017 Int. Conf. and Workshop on Bioinspired Intelligence, pp. 1–4. IEEE, Funchal, Portugal (2017).
19. Wu, H., Soragan, J., Lowit, A., Di-Caterina, G.: A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks. In: 19th Annual Conf. Int. Speech Communication Association INTERSPEECH 2018, pp. 446-450. ISCA, Hyderabad, India (2018).
20. Alhussein, M., Muhammad, G.: Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access (6), 41034–41041 (2018).
21. Roy, S., Sayim, M. I., Akhand, M. A. H.: Pathological Voice Classification Using Deep Learning. In: 1st Int. Conf. Advances in Science, Engineering and Robotics Technology, pp. 1-6. IEEE, Dhaka, Bangladesh (2019).