Convolutional Neural Networks for Pathological Voice Detection

Huiyi Wu, John Soraghan *Senior Member, IEEE*, Anja Lowit, Gaetano Di Caterina

***Abstract* — Acoustic analysis using signal processing tools can be used to extract voice features to distinguish whether a voice is pathological or healthy. The proposed work uses spectrogram of voice recordings from a voice database as the input to a Convolutional Neural Network (CNN) for automatic feature extraction and classification of disordered and normal voice. The novel classifier achieved 88.5%, 66.2% and 77.0% accuracy on training, validation and testing data set respectively on 482 normal and 482 organic dysphonia speech files. It reveals that the proposed novel algorithm on the Saarbruecken Voice Database can effectively been used for screening pathological voice recordings.**

I. Introduction

Dysphonia is the global term used for disorders of voice production, either due to structural changes in the larynx or functional/behavioral issues. Patients diagnosed with dysphonia often present with prolonged hoarseness or possibly loss of voice. Figures estimate that around 10% of the overall working population, and between 3%-9% of the overall adult population might experience some problems with their voice at any given time [1]. Clinical assessment tends to be based on perceptual analysis of voice. However, this is highly subjective and outcomes can vary depending on the clinician's level of training and experience with dysphonia.

Acoustic analysis of pathological voice detection has been popular to supplement perceptual analysis as it is non-invasive and provides robust quantitative measures. In this case, acoustic analysis becomes a popular alternative tool for pathological voice diagnosis in the recent years. In general, signal processing tools are applied to find the proper feature set, and machine learning methods are used for dimensionality reduction and classification to diagnose dysphonia.

Deep Learning in Machine Learning field has become a powerful classification framework demonstrating superior performance in many application domains such as computer vision and speech recognition. However, to date very little publications in the use of deep learning technologies for pathological voice analysis have been reported.

In this paper, we propose a method Convolutional Neural Network (CNN), which originated from deep learning field, to extract features automatically from the spectrogram of voice recordings for dysphonia diagnosis. Small windows are used as feature extractors for the spectrogram to detect the subtle variations between pathological voice and normal voice which are hard to be manually detected using conventional method.

The rest of the paper is organized as follows. Section II introduces the related works in acoustic measurement for pathological voice detection. Section III explains the details of the methodology and process of the experiments. In Section IV, results are represented. Section V provides a conclusion on the results reported, and proposes ideas for future works.

II. Related Work

There is a large amount of related works exist. Many approaches [2-6] extract signal processing features such as Mel-frequency Cepstral Coefficients (MFCC), Wavelet Packet Transform (WPT), while some uses multidimensional voice program (MDVP) parameters according to physiological and etiological reasons. MDVP parameters including pitch, jitter and shimmer are used to detect the roughness of the speech, while others such as Harmonic-to-Noise Ratio (HNR), Normalized noise Energy (NNE) and Glottal-to-Noise Ratio (GNR) represent the breathiness of the speech. Furthermore, dimensionality reduction methods such as Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA), kernel PCA, Fisher Discriminant Ratio (FDR) and Singular Value Decomposition (SVD) etc. are used for searching the suitable latent variables for classification. k- Nearest Neighbor (kNN), Random Forests (RF), Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Neural Networks (NN) are applied for classification.

TABLE I. Overview of Related Works with MEEI Database

	Data Source[a]	Feature Set	Feature selection	Classifier	Accuracy
[2]	710-53	MFCC, pitch	-	HMM	98.59%
[3]	67-53	WPT	LDA, PCA	SVM	100%
[4]	657-53	WPT	SVD	k-NN	100%
[5]	657-53	NLD	-	GMM, SVM	98.23%
[6]	53-95	MDVP	FDR	SVM	88.21%

a. Data amount (Pathological-Normal)

As indicated in Table 1 many research works has been carried out using the MEEI database. However, these works reveal "perfect" result that leads to researchers to question the usefulness of the database. Muhammad et al. in [7] explains that this is because the normal and pathological voice recordings are recorded in two different environments in this

*Research supported by Capita plc.

Huiyi Wu (email: huiyi.wu@strath.ac.uk), John Soraghan, Gaetano Di Caterina are with the Centre for Signal and Image Processing, Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK.

Anja Lowit is with Speech and Language Therapy, School of Psychological Sciences and Health, University of Strathclyde, Glasgow, G1 1QE, UK

database. Therefore, it is hard to distinguish whether the system is classifying voice features or environments. In addition, due to the unbalanced data structure, with 11 times more data from pathological to normal files [2, 4, 5], the classification result is effected to some extent.

Saarbruecken Voice Database is a database with recordings that are all sampled at 50 kHz and with 16-bit resolution, which makes it a reliable database for research. Because this is a new database, little research has been done using this database, which are listed in Table II.

TABLE II. OVERVIEW OF RELATED WORKS WITH SV DATABASE

	Data Source[c]	Feature Set	Dimensionality Reduction and Classifier	Accuracy
[6]	262-244	MDVP	FDR, SVM	99.68% [b]
[8]	266-263	AC	-	98.94% [b]
[9]	255-255	MDVP, MFCC	kPCA, RF	100% [a]
[10]	1320-650	MFCC, HNR, NNR, GNR	GMM	79.40% [a] 67.00% [b]
[11]	480-480	-	DNN	68.08% [b]

a. Accuracy with fusion of sustained vowels /a/, /i/ and /u/
b. Accuracy with sustained vowel /a/
c. Data amount (Pathological-Normal)

From the literatures, we can see that Saarbruecken Voice Database appears more challenging while more trust-worthy for experiments. Some experiments use small amount of data and achieved almost 100% accuracy using statistical methods [6, 8, 9]. This is questionable compared to [10] using GMM-HMM which achieves 67.00% accuracy when the data amount is large. In [11], Deep Learning has been used for the first time, applying Long Short-Term Memory (LSTM), a type of recurrent neural network and using information from the time-domain axis. However, since pathological voice contains information without regard to time, this model might not be the most proper one for this problem.

III. METHODOLOGY

A. Data source

Saarbruecken Voice Database is a German database with a collection of voice recordings from more than 2000 individuals, and it is collected by the Institute of Phonetics of Saarland University. Each participant file contains recordings of sustained vowels /a/, /i/ and /u/ in low, neutral, high and low-high-low pitch and a continuous speech sentence “Guten Morgen, wie geht as Ihnen?” (“Good morning, how are you?”). All recordings are recorded in 50 kHz sampling frequency and 16-bit resolution. It is proved to be superior to MEEI database because it is recorded in the same environment[7].

Saarbruecken voice database contains 71 different pathologies. Some pathologies belong to functional dysphonia type, including hyper-functional dysphonia, hypo-functional dysphonia and psychogenic dysphonia. Other pathologies are mostly organic dysphonia which is caused by structural changes in the vocal cord. This type of dysphonia contains significant characteristics to be detected so that it is chosen for this experiment. We select 6 pathologies (Laryngitis, leukoplakia, Reinke’s edema, recurrent laryngeal nerve paralysis, vocal fold carcinoma, vocal fold polyps) as the pathological group. We use sustained vowel /a/ at neutral pitch of each individual, of which 482 were healthy and 482 are diagnosed with pathologies (140 laryngitis, 41 leukoplakia, 68 Reinke’s edema, 213 recurrent laryngeal nerve paralysis, 22 vocal fold carcinoma and 45 vocal fold polyps) (some pathologies repeat in the same file).

The data is divided into training set and testing set with 75% and 25% samples respectively. Therefore, there are 724 training and 240 testing files in all.

B. Pre-processing for Input Data to CNN

To use CNN for application, a 2-Dimensional graph is ideal for extracting features. In this case, we need to perform some pre-processing steps to form the feature map to feed into the CNN system.

The procedure is shown in Figure 1. For implementation, the Python programming language has been used with the signal processing package *scipy.signal*. The original speech is first resampled at 25 kHz. Furthermore, Short-Time Fourier Transform (STFT) are applied to the resampled data for transforming the time-domain signal into spectral-domain signal. Compared to time-domain representation, spectral-domain signals contain more pathological information. Some research works demonstrated that pitch, formants and NHR, HHR, GNR etc. can represent some characteristics of the pathological voice such as hoarseness, breathiness and roughness[12, 13], which can all be seen on spectrograms. In STFT, each file use 10 ms *hamming* window segments, with 50% overlap between consecutive windows. Finally, the spectrograms are reshaped to the same size of 60*155 points, with 155 being the minimum length of the spectrograms. This is because there is a large part of the area in the spectrogram contains no information, and the useless part of the spectrogram is cut off to reduce the effect of noise to the classification result. The comparison of the spectrograms are shown in Figure 2.

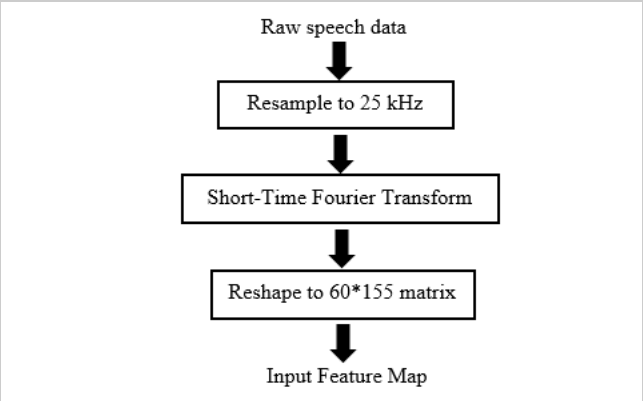

Figure 1. Pre-processing steps for Input data

C. CNN Architecture

Pathological voice contains subtle differences that can be seen on the spectrogram compared to normal voice, which are difficult to be manually defined using particular criteria. Hence the CNN plays an important role as a feature extractor

to distinguish two classes using specific features. The CNN architecture is shown in Figure 3.

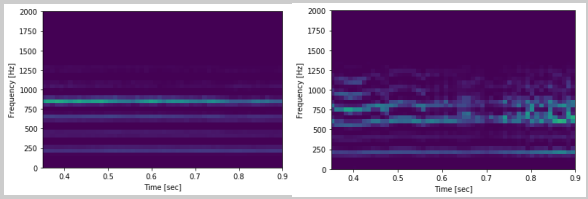

Figure 2. Comparison of input feature map *(a).spectrogram of one normal voice; (b). spectrogram of one pathological voice*

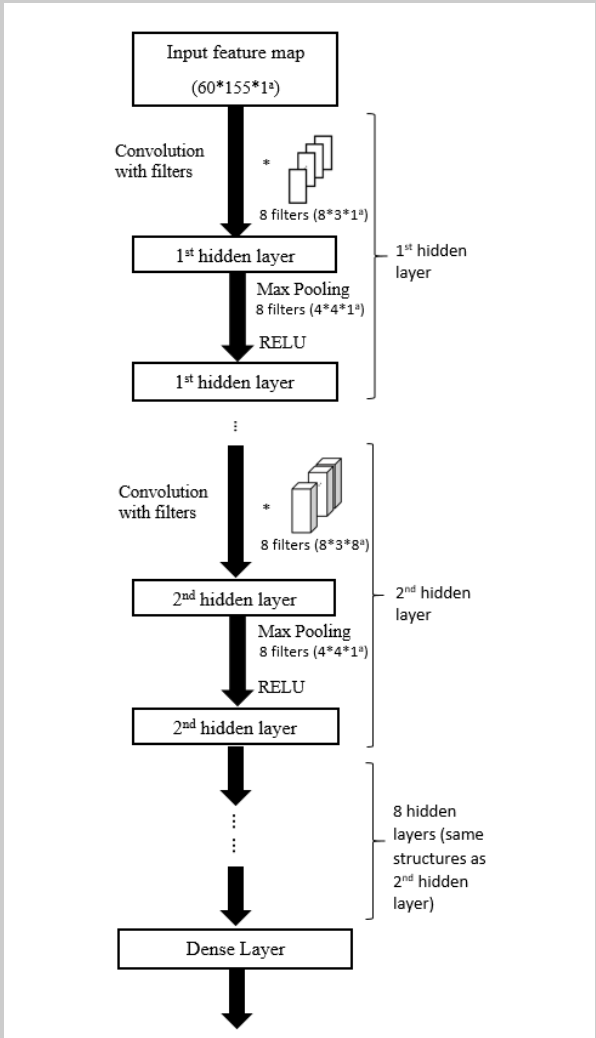

Figure 3. CNN architecture *(a. length*width*depth)*

The size of the input feature map is 60*155*1. Since it is the spectrogram of the speech file, the depth of this input layer is 1, and it has the same meaning as "color channels", i.e. Red-Green-Blue (RGB) in computer vision field; in other words, the spectrogram can be seen as a grey scale image.

The input feature map is then convolved with a set of 8 filters. Each filter has the shape of 8*3*1 and stride of 1. We use the rectangular filters in this work due to the spectrogram characteristics. Furthermore, max-pooling filters with the shape 4*4 and stride of 1 are applied to pool the significant values out and reduce the computational complexity. Then the activation function RELU is applied to make the neural network non-linear and fit for classification.

After the first hidden layer, each layer was convolved with 8 filters with the shape 8*3*8 and stride of 1. Max-pooling filters and activation function is the same as for the first hidden layer.

After 10 hidden layers to extract the features from the spectrogram, the feature map is formed into a Dense Layer, which is a fully-connected layer, to train the model for classification. L2-regularization is used in this layer to avoid overfitting problems.

D. Hyper-parameter Setting

Python based Tensorflow[14] is used as a framework for the training process. Because the mini-batch gradient descent use GPU for matrix computation and will lead to high speed, the training samples are divided into 256 samples in each mini-batch to be trained on GPU NVidia GTX1070 in this work. Adam Optimizer[15] is applied in this experiment with initial learning rate 0.0006 so that the training process becomes more robust. Delta value of the L2 regularization is set to 0.0001 and the maximum epochs of training is 100.

IV. Experimental Results

Confusion matrix of validation dataset and testing dataset are listed respectively in Table III and Table IV. Several metrics indicating the classification result are shown in Table V. Sensitivity (SN) reveals how good the classifier is at detecting the pathological voice files, which has the same meaning as "recall" and is calculated as in (1). Specificity (SP) calculated as in (1) reveals the proportion of normal voice files that are correctly identified. Precision (P) shows how many of the pathological voice files classified are relevant, and F1-score (F1) has also been taken into account, calculated as in (3).

$$SN = \frac{TP}{TP + FN}, SP = \frac{TN}{FP + TN} \quad (1)$$

$$P = \frac{TP}{TP + FP}, F1 = 2\frac{P \cdot SN}{P + SN} \quad (2)$$

True Negative (TN) represent normal voice recordings that are correctly detected as "normal voice"; True Positive (TP) represent pathological voice recordings that are correctly detected as "pathological voice"; False Negative (FN) represent pathological voice recordings that are detected as "normal voice", False Positive (FP) represent normal voice recordings that are classified as "pathological voice".

It can be seen from Table V that the classifier achieved overall accuracy (ACC) of 88%, 66% and 77% on training dataset, validation dataset and testing dataset respectively. Compared to [11], spectrogram features show greater performance on pathological voice detection than raw time-domain signals. Moreover, the proposed algorithm is shown to be more robust for dealing with large amount of data compared to [6, 8, 9]. However, training data accuracy

achieved much better than validation set and testing set, which reveals overfitting problem to some extent.

TABLE III. CONFUSION MATRIX OF VALIDATION DATASET

	True: pathological	**True: normal**
Prediction: pathological	39	25
Prediction: normal	24	57

TABLE IV. CONFUSION MATRIX OF TESTING DATASET

	True: pathological	**True: normal**
Prediction: pathological	61	19
Prediction: normal	14	51

TABLE V. METRICS TO MEASURE THE CLASSIFIER

Dataset	**Metrics**				
	SN(r)	***SP***	***p***	***F1***	***ACC***
Training dataset	0.93	0.83	0.85	0.89	0.88
Validation dataset	0.61	0.70	0.61	0.61	0.66
Testing dataset	0.76	0.79	0.81	0.78	0.77

V. CONCLUSION

Our results have shown that spectrograms can been used effectively as the input to classify pathological voice and normal voice, without the necessity to extract features manually. In this work, organic dysphonia was selected as the pathological group, as it shows more significant pathological characteristics than functional dysphonia. However, the high accuracy on training dataset compared to the validation and testing dataset reveals an overfitting phenomenon on the classifier, which is an old *Bias and Variance* dilemma in deep learning field. Different CNN structure has been changed to eliminate the problem. For example, reducing the number of nodes of CNN, increasing the size of the filters, performing *drop-out* on convolutional layers, and adding L2 regularization on Dense Layer. We conduct hundreds of experiments to choose the most appropriate parameters and structures for CNN. However, the best way might be using larger amounts of training data, while requiring spending time and resources to collect.

In the future work, more data will be collected with Glasgow Royal Infirmary. At the same time, subtle characteristic differences between functional dysphonia and organic dysphonia will be investigated and compared together with normal voice recordings. Deep learning techniques and traditional data mining tools will be compared to explore a better approach for dysphonia screening.

ACKNOWLEDGMENT

The authors would like to acknowledge Capita plc and University of Strathclyde for their financial support with this study.

REFERENCES

[1] K. Verdolini and L. O. Ramig, "Occupational risks for voice problems," *Logopedics Phoniatrics Vocology,* vol. 26, pp. 37-46, 2001.

[2] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, 2002, pp. 182-183 vol.1.

[3] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomedical Signal Processing and Control,* vol. 7, pp. 3-19, 2012/01/01/ 2012.

[4] M. Hariharan, K. Polat, and S. Yaacob, "A new feature constituting approach to detection of vocal fold pathology," *International Journal of Systems Science,* vol. 45, pp. 1622-1634, 2014/08/03 2014.

[5] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 370-379, 2011.

[6] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, *et al.*, "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification," *Journal of Voice,* vol. 31, pp. 113.e9-113.e18.

[7] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, *et al.*, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical Signal Processing and Control,* vol. 31, pp. 156-164, 2017/01/01/ 2017.

[8] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions," *Journal of Voice,* vol. 31, pp. 3-15.

[9] D. Hemmerling, A. Skalski, and J. Gajda, "Voice data mining for laryngeal pathology assessment," *Computers in Biology and Medicine,* vol. 69, pp. 270-276, 2016/02/01/ 2016.

[10] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," in *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, *et al.*, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99-109.

[11] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice Pathology Detection Using Deep Learning: a Preliminary Study," in *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, 2017, pp. 1-4.

[12] J. Rafael Orozco Arroyave, J. Francisco Vargas Bonilla, and E. Delgado Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents on Signal Processing,* vol. 2, pp. 96-107, 2012.

[13] A. Akbari and M. K. Arjmandi, "An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features," *Biomedical Signal Processing and Control,* vol. 10, pp. 209-223, 2014/03/01/ 2014.

[14] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467,* 2016.

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