Deep Learning Approaches for Pathological Voice Detection Using Heterogeneous Parameters

JiYeoun LEE†, Member and Hee-Jin CHOI††, Nonmember

SUMMARY We propose a deep learning-based model for classifying pathological voices using a convolutional neural network and a feedforward neural network. The model uses combinations of heterogeneous parameters, including mel-frequency cepstral coefficients, linear predictive cepstral coefficients and higher-order statistics. We validate the accuracy of this model using the Massachusetts Eye and Ear Infirmary (MEEI) voice disorder database and the Saarbruecken Voice Database (SVD). Our model achieved an accuracy of 99.3% for MEEI and 75.18% for SVD. This model achieved an accuracy that is 7.18% higher than that of competitive models in previous studies.

Table 1 Overview of related works

| Article | Feature set and Database | Classifier | Accuracy (%) |
|---------|--------------------------|------------|--------------|
| [5]     | MFCC, sustained vowel, MEEI | DNN        | 99.14        |
| [4]     | MFCC, sustained vowel, MEEI | GMM, SVM   | 96.1         |
| [9]     | Spectrogram, sustained vowel, SVD, MEEI | CNN(VGG16 net, CaffeNet) | 93.9 |
| [1]     | MFCC, HOS, sustained vowel, MEEI | GMM, LDA, CART, | 92.7 |
| [7]     | MFCC, sustained vowel, SVD | ANN        | 87.82        |
| [8]     | STFT, sustained vowel, SVD | CNN, CBNN  | 68           |
| [6]     | STFT, sustained vowel, SVD | CNN        | 66.2         |

1. Introduction

Automatic detection of disordered voice has attracted considerable clinical and academic interest in the hope of accurately diagnosing voice disorders before confirmation by well-trained specialists and expensive equipment. Although many researchers have focused on the automatic detection of voice pathologies using acoustic analysis, parametric and non-parametric feature extraction, pattern recognition algorithms, and statistical methods [1]–[4], research in the area of voice disorder detection using deep learning techniques recently been published.

Table 1 shows the overview of related works. The deep neural network (DNN) algorithm could fully utilize the acoustic features and efficiently differentiate between normal and pathological voice samples [5]. A convolutional neural network (CNN) and short-time Fourier transform (STFT) were used for the reliable classification and feature detection of voice pathologies [6]. Mel-frequency cepstral coefficients (MFCC) derived from the Saarbruecken voice database (SVD) are analyzed using an artificial neural network (ANN) and a support vector machine (SVM). Recently, a method for classifying healthy and pathological voices using ANNs has been proposed [7]. A convolutional deep belief network (CDBN) that uses spectrograms of normal and pathological speech as the input has been developed. Accuracies of the validation and testing set using the SVD are 68% and 71% respectively [8]. A voice pathology detection system was embedded in a mobile multimedia healthcare framework using deep learning to constantly assess the voice condition of a patient [9].

Despite the success of the above-mentioned studies, automatic expert detection of voice disorders is not widely used because people are still concerned about the use of machines for diagnosing voice disorders. In this paper, we propose a new approach for enhancing the performance of pathological voice classification using the combination of a CNN, feedforward neural network (FNN), and various parameters such as MFCCs, linear predictive cepstral coefficients (LPCCs), and higher-order statistics (HOS). No study has hitherto attempted to use HOS with deep learning methods for the detection of pathological voice samples. Most of the research works use the Massachusetts Eye and Ear Infirmary (MEEI) database. Although healthy and pathological voices in this database were recorded in two different environments [8], it has been used for comparing performance of our study with that of others [1]–[3]. The SVD is relatively new; therefore, little research has been carried out using it. The SVD is a downloadable database with all recordings sampled at 50 kHz with a 16-bit resolution. The audio samples are recorded in the same environment; therefore, it is an ideal database for this research.

Our study was conducted with the following objectives: (1) to propose a FNN and CNN based system for detecting MFCC, LPCC, and HOS features extracted from voice samples, (2) to examine the performance of FNN and CNN in differentiating between normal and pathological voice samples, and (3) to validate the accuracy of the FNN and CNN using the widely applied MEEI and SVD voice disorder databases.
2. Methodology

2.1 Database

Vocal signals were collected from the MEEI Voice Disorders Database [10]. The sustained /a/ sounds (1–3 s) phonated by 53 normal speakers and 173 pathological speakers with a wide range of organic, neurological, traumatic, and psychogenic voice disorders were selected. The extracted subset was chosen to be the same as the one used in previous studies [1]–[4] to enable the comparison of our results with those of previous studies.

This work also uses the SVD, which was recorded by the Institute of Phonetics of the Saarland University in Germany [11]. We use the sustained vowel /a/ sound recorded from each individual at neutral pitch, of which 482 are healthy and 482 are diagnosed with various pathologies (140 laryngitis, 41 leukoplakia, 68 Reinke’s edema, 213 recurrent laryngeal nerve paralyses, 22 vocal fold carcinoma, and 45 vocal fold polyps) [6], [8]. The extracted subset is the same as that described in previous studies [6], [8] to enable the comparison of our results with those of the aforementioned previous studies.

2.2 Deep Learning Methods

In the FNN, as shown in Fig. 1, the information moves in only one direction, that is, forward, from the input nodes through the hidden nodes and to the output nodes. There are no cycles or loops in the network.

We use the FNN to solve the binary classification problem of the classification of normal and pathological voices. In machine learning, classification is a supervised learning method where the task is to divide the data samples into predefined groups using a decision function. We use two feed forward layers. The first layer is followed by rectified linear unit (ReLU) activation, and the last layer is followed by Softmax activation.

CNN is very similar to ordinary neural networks, as shown in Fig. 2. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product, and optionally follows it with a non-linearity.

The whole network still expresses a single differentiable score function, that is, from the raw image pixels on one end to the class scores at the other end. It still has an activation function (e.g. SVM/Softmax) on the last fully connected layer.

We use two convolutional layers and two feed forward layers. Dropout with a probability of 0.05 and batch normalization were applied to all layers except the feed forward layers. Max-pooling and average-pooling operations are performed between the convolutional layers to downsample the intermediate representations in time and to add some time invariance in the process. We used a filter size of eight frames and stride one in every convolutional layer with max-pooling at a pool size of eight and stride eight between the first and second convolutional layers, and average-pooling at a pool size of eight and stride two between the other layers. The output of the last fully connected layer with size 64 is fed to a 2-way Softmax classifier, which classifies a patch into one of the two classes.

2.3 Proposed Parameter Combinations

Among various HOS, the 3rd- and 4th-order cumulants called the normalized skewness, $\gamma_3$, and the normalized kurtosis, $\gamma_4$, are used as characteristic parameters in this study. They can be defined as shown in (1).

$$\gamma_3 = \frac{\sum_{n=1}^{N}(x_n - \mu)^3}{(N-1)\sigma^3}, \quad \gamma_4 = \frac{\sum_{n=1}^{N}(x_n - \mu)^4}{(N-1)\sigma^4}$$

(1)

Where $x_n$ is the $n^{th}$ sample value and $N$ is the number of the samples, while $\mu$ and $\sigma$ represent the mean and the standard deviation, respectively.

Various parameter combinations are suggested using three kinds of parameters to investigate classification performance. For example, Eq. (2) shows the parameter vectors...
consisting of the MFCCs, LPCCs, $\gamma_3$, and $\gamma_4$ extracted by each analysis frame. Based on Eq. (2), our experiments are carried out with parameter combinations created by adding and subtracting each parameter.

\[
\vec{p} = [MFCC_1, \ldots, MFCC_k, LPCC_1, \ldots, LPCC_k, \gamma_3, \gamma_4]
\]

where $MFCC_k$ and $LPCC_k$ are the $k^{th}$ coefficients, respectively.

3. Experiments and Results

Figure 3 (a) and (b) present the normalized skewness and kurtosis in the form of box plots to provide better visualizations of normal and pathological voice signals. As may be seen from Table 2, the normalized skewnesses estimated in pathological voices tend to be lesser than zero. In case of the normalized kurtosis, one estimated in pathological signals tend to be larger than one estimated in normal voice signals. The statistical analyses between normal and pathological voice signals were carried out using Mann-Whitney’s U test for independent samples. The significance level was set a priori at $(P < 0.05)$. The Mann-Whitney U test showed a significant difference between the normal and pathological voice signals regarding the normalized skewness $(P = 0.000)$ and kurtosis $(P = 0.000)$. It is evident that they are useful and meaningful for the classification of pathological voice signals.

The framework for the training process was developed in Python using PyTorch\[12\]. In the training step, FNN and CNN are trained using various parameter combinations extracted from the pathological and normal voices, respectively.

In Tables 3 and 4, the first and second parameters, made up of 20-dimensional MFCCs and LPCCs, were extracted from a 40-millisecond window signal using a 20-millisecond frameshift, respectively. The same settings were applied in many previous studies\[12\]. The next parameter, MFCC+HOS, was created by appending two parameters to the original 20-dimensional MFCCs, and thus, it had 22 dimensions. The LPCC+HOS parameter also has the same dimensions as that of MFCC+HOS. MFCC+LPCC parameter has 40-dimensions. Finally, the MFCC+LPCC+HOS parameter, which is made up of 20-dimensional MFCCs, 20-dimensional LPCCs, skewness, and kurtosis, had 42 dimensions. Therefore, the dimension of $\vec{p}$ shown in Eq. (2) ranges from 20 to 42.

All voice data were grouped into training (70% of the data) and test (30%) sets to implement all methods. Each set for a fivefold cross validation scheme was randomly selected from the subset\[1\]–[4].

Table 3 and Table 4 show the final experimental results for combinations of various parameters with FNN and CNN using MEEI and SVD, respectively. MFCC as a baseline parameter showed better performance than LPCC. The proposed parameters also showed better accuracy than the baseline parameters. Overall, the performance of the FNN was considerably better than that of the CNN using the SVD. The performance of the CNN was better than that of the FNN when using the MEEI database. In the case of using the MEEI, various combinations, such as MFCC+HOS, MFCC+LPCC, and MFCC+LPCC+HOS with FNN and
CNN, and LPCC+HOS with CNN, achieved the best performance at 99.3%. In the case of using the SVD, the proposed combinations of MFCC+LPCC+HOS with CNN and MFCC+LPCC with CNN achieved the best performances, at 74.82% and 75.18%, respectively. The greater the number of parameters that were used in the two deep learning methods, better the performance.

The best performance among the FNNs was the FNN that used a combination of the MFCC, LPCC, and HOS parameters. The best among the CNNs was the one that used a combination of the MFCC and LPCC parameters. The relative difference of the best performance of FNN and CNN was 0% and 0.35%, respectively. As two different parameters were added to the baseline parameter, such as MFCC or LPCC, their effectiveness was evaluated through experiments. As shown in Table 1, the model used by Fang et al. [5] achieved the highest accuracy, 99.14%, using the MEEI database among the previous models. Our proposed model also achieved relatively high performance, 99.3%, compared with that of Fang et al.’ model. In addition, the research conducted by Huiyi et al. [8] achieved the highest accuracy using the SVD compared with other studies as shown in Table 1. However, our proposed model achieved 6.82% higher performance than that of the model proposed by Huiyi et al. Although Souissi’s study demonstrates high performance using the SVD, the number of databases (50 normal and 70 pathological voices) used is too small compared with that of our study (482 normal and 482 pathological voices). As shown in Tables 3 and 4, there was a performance difference of up to about 14% depending on the number of database used. Also, we didn’t know what kind of data was selected and tested in SVD.

4. Conclusion

In this study, novel deep learning models such as FNN and CNN were presented to combine various parameter representations for pathological voice detection. In several studies, MFCC has been used as the significant parameter for classifying pathological and normal voices [1], [4]–[9]. It is necessary to increase performance through a combination of heterogeneous parameters; therefore, we introduced some effective combinations of heterogeneous parameters in this study. The HOS parameters were useful for an analysis between normal and pathological voices because they showed p-values < .05. The proposed FNN model using the combination of MFCC, LPCC and HOS parameters obtained the highest accuracy of 99.3% for the MEEI database and 75.18% for the SVD. The FNN with heterogeneous parameters such as MFCC, LPCC, and HOS might be usefully applied to complement existing pathological voice detection methods in the clinic. We will continue to study features that reflect other information about pathological voices in order to realize higher performances in pathological voice detection of the real-world environment.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) (No. 2017R1A2B4011373).

References

[1] J.-Y. Lee, S. Jeong, and M. Hahn, “Pathological Voice Detection Using Efficient Combination of Heterogeneous Features,” IEICE Trans. Inf. & Syst., Vol.E91-D, no.2, pp.367–370, Feb. 2008. DOI: http://dx.doi.org/10.1093/ietisy/e91-d.2.367.
[2] J.-Y. Lee, S. Jeong, H.-S. Choi, and M. Hahn, “Objective Pathological Voice Quality Assessment Based on HOS Features,” IEICE Trans. Inf. & Syst., Vol.E91-D, no.12, pp.2888–2891, Dec. 2008. DOI: http://dx.doi.org/10.1093/ietisy/e91-d.12.2888.
[3] J.Y. Lee and M.S. Hahn, “Automatic assessment of pathological voice quality using higher-order statistics in the LPC residual domain,” EURASIP J. Adv. Signal Process., Vol.5, no.1, pp.1–8, 2010. DOI: http://dx.doi.org/10.1155/2009/748207.
[4] X. Wang, J. Zhang, and Y. Yan, “Discrimination between pathological and normal voices using GMM-SVM approach,” Journal of Voice, Vol.25, no.1, pp.38–43, Jan. 2011. DOI: http://dx.doi.org/10.1016/j.jvoice.2009.08.002.
[5] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, C.-T. Wang, “Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach,” J. Voice, In press, 2018. https://doi.org/10.1016/j.jvoice.2018.02.003.
[6] H. Wu, J. Soraghan, A. Lowit, and G.D. Caterina “Convolutional Neural Networks for Pathological Voice Detection,” Conf. Proc. IEEE Eng. Med. Biol. Soc., pp.1–4, 2018. DOI: 10.1109/EMBC.2018.8513222.
[7] N. Souissi and A. Cherif, “Artificial Neural Networks and Support Vector Machine for Voice Disorders Identification,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol.7, no.5, 2016. http://dx.doi.org/10.14569/IJACSA.2016.070546.
[8] H. Wu, J. Soraghan, A. Lowit, G. Di-Caterina, “A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks,” pp.446–450, Sep. 2018. 10.21437/Interspeech.2018-1351.
[9] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, “Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System,” Mobile Information Systems, Vol.2016, Article ID 7805217, 12 pages, 2016. https://doi.org/10.1155/2016/7805217.
[10] Kay Elemetrics Corp., Disordered Voice Database (CD-ROM), Version 1.03, Massachusetts Eye and Ear Infirmary (MEEI), Voice and Speech Lab., Boston, MA, 1994.
[11] J.B. William and P. Manfred, “Saarbrucken voice database,” Institute of Phonetics, Univ. of Saarland, 2007. http://www.stimmdatenbank.uni-saarland.de/.
[12] P. Adam, G. Sam, C. Soumith, and C. Gregory, “Automatic differentiation in PyTorch,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.