CEUR-WS.org/Vol-2893/short_5.pdf

Detection of Defective Speech Using Convolutional Neural Networks

Mikhail Belenko, Nikita Burym and Pavel Balakshin

ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russian Federation

Abstract

This paper presents an algorithm for detecting a pathological voice. It is shown that the convolutional neural network effectively extracts features from the spectrograms of voice recordings and diagnoses voice disorders. The deep belief convolutional network helps to initialize weights and makes the system more reliable. The effect of the size of convolutional network filters on each layer on the system performance is also studied.

Keywords

Speech recognition, Defective speech, Convolutional Neural Network, Convolutional Deep Belief Network.

1. Introduction

Automatic detection of pathological voice disorders, such as paralysis of the vocal cords or Reinke's edema, is a complex and important problem of medical classification. While deep learning methods have made significant progress in speech recognition, fewer studies have been conducted in the detection of pathological voice disorders. This paper presents a new system of pathological voice recognition using convolutional neural network (CNN) as the basic architecture. The new system uses spectrograms of normal and abnormal speech recordings as input to the network. Initially, the deep belief convolutional network (CDBN) is used to pretrain CNN weights. It acts as a generative model for studying the structure of input data using statistical methods. CNN then uses training with controlled back propagation to adjust the weights. As a result, it is clear that a small amount of data can be used to achieve good results in classification using this approach. The performance analysis of this method is performed using real data from the SaarbruckenVoice database.

Voice pathologies affect the larynx and lead to irregular fluctuations in the vocal folds. This leads to psychological and physiological problems for individuals, and also has a significant impact on the economy, taking into account the costs of medical diagnosis and treatment. The traditional method of diagnosing voice pathology relies on the experience of a doctor and on expensive devices such as a laryngoscope, endoscope, etc. However, computer-based medical systems for diagnosing voice pathologies are becoming popular due to significant advances

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10–11, 2020, Online Saint Petersburg, Russia

0000-0002-5060-1512 (M. Belenko); 0000-0002-4343-6408 (N. Burym); 0000-0003-1916-9546 (P. Balakshin)

© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

in signal processing technologies. These comprehensive tools are usually non-invasive and non-subjective, which is generally an advantage in the medical field[1].

Over the past few decades, many scientific works have been carried out related to the automatic detection of voice pathologies. Usually, these features are extracted from speech recordings and then processed by classifiers to distinguish normal speech from pathological speech. Signs are mainly derived from two areas of research. One of them is related to speech recognition applications, where signal processing tools are used to automatically detect signal properties such as Mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and the energy and entropy of discrete wavelet packets[2-4].

Other signs come from measuring voice quality in accordance with physiological and etiological studies. While pitch, jitter, and flicker are used to determine the depth of speech, other characteristics such as harmonic-to-noise ratio (HNR), normalized noise energy (NNE), laryngeal-to-noise ratio (LNR), and cepstral peak prominence (CPP) represent speech hoarseness[5]. Most research papers use the Massachusetts Eye and year Infirmary (MEEI) database.However, healthy voice recordings and abnormal voice recordings in this database are recorded in two different environments[6], which makes it difficult to distinguish whether these are discriminating environments or voice features.The Saarbruecken Voice Database is a downloadable database with all recordings sampled at 50 kHz and 16-bit resolution. This database is relatively new, So little research has been done on it. However, the recordings are recorded in the same environment, so it was decided to choose it for this study.

Modern signal processing techniques previously used in the field of speech recognition have also made significant progress in the field of automatic detection of abnormal voice. For example, in [7], the Russian language Gaussian Mixture Model (GMM) based on the Saarbruecken voice database is used, and 67% classification accuracy is achieved with a neutral stable vowel /a/. However, with the increasing computing capabilities of hardware and the improvement of machine learning algorithms, the Markov model hidden in the deep neural network gradually replaces the traditional GMM-HMM [8] and becomes a popular method of speech recognition. To date, deep learning methods are not commonly used in the field of pathological voice detection, mainly due to the limited amount of data, since DNN requires a large amount of data for training. In [9], a restricted Boltzmann machine (RBM) is proposed as an unsupervised method for pre-training DNN to accurately achieve global minima. As a generative model, it improves deep learning performance even on small datasets. Deep belief convolutional networks (CDBNS) were proposed in [10] as an advanced specific structure for CNN pre-training. This article considers a new deep learning method for automatic detection of abnormal voice. In this paper, we use the CNN convolutional neural network structure for automatic analysis of speech recording spectrograms. CDBN is used for pre-training weights and preventing problems with over-training. A similar approach is proposed in [11], but the influence of convolutional neural network parameters is left behind in that study.

2. Methodology

Figure 1 shows a block diagram of the proposed system for detecting abnormal voice. First, preprocessing is applied to speech recordings, which includes resampling and shape-changing

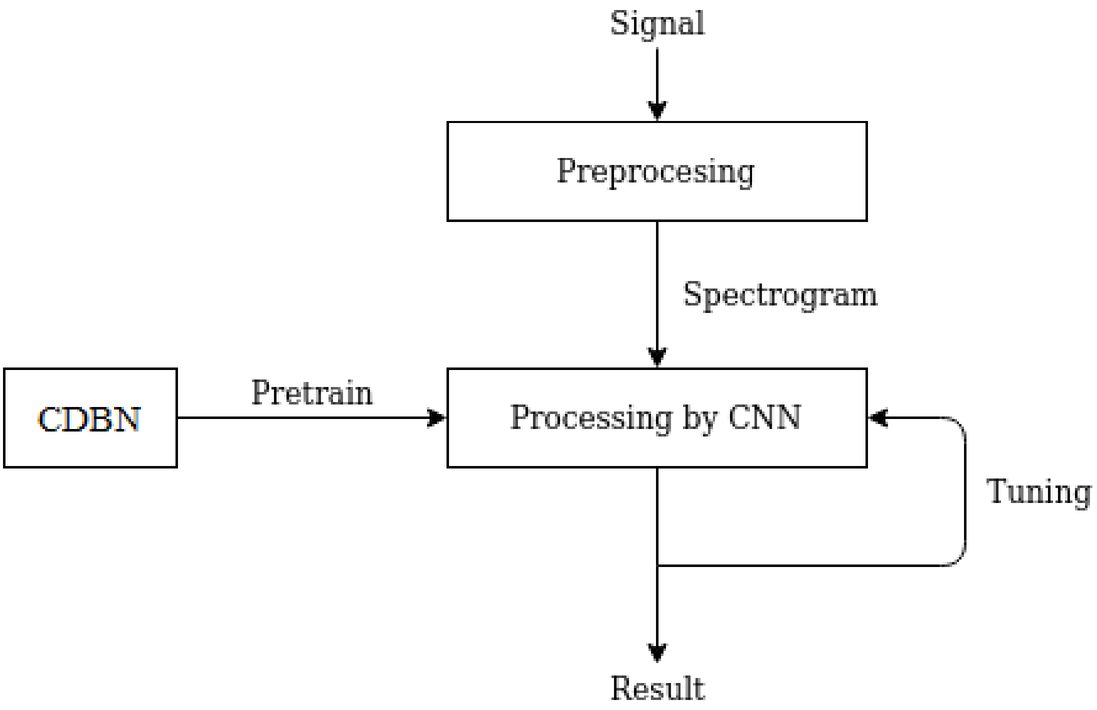

Figure 1: System architecture.

methods. Then a short-time Fourier transform (STFT) is applied to obtain speech recording spectrograms as input to the CNN system. Weights in the CNN system are pre-trained using CDBN and adjusted using the back propagation method. The trained CNN system is able to automatically extract features and classify audio samples.

2.1. Input data

One of the properties of CNN is the ability to reduce the dimension of two-dimensional feature maps. Therefore, speech recordings are converted from one-dimensional signals to two-dimensional spectrograms.

2.1.1. Dataset

This paper uses the Saarbruecken voice database, which was registered by the Institute of phonetics of the Saarland University in Germany. This database contains 71 different pathologies with speech recordings from more than 2000 people. Each participant's file contains recordings of the sustained vowels /a/, /i/, and /u/ inneutral, low, high, and low-high-low intonations, and the continuous speech sentence "Guten Morgen, wie geht as Ihnen?" ("Good morning, How are you?"). Stable vowels are used in this work because they are stationary in time and it is easier to see changes.

The following pathologies were selected as the pathological group

- laryngitis
- leukoplakia
- Reinke's edema
- paralysis of the recurrent laryngeal nerve
- carcinoma of the vocal folds
- polyps of the vocal fold.

All these pathologies are organic dysphonia, which are caused by structural changes in the vocal cord. The vowel /a/ is used at a neutral height for each individual, of which 482 are 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).The data is divided into a training set and a test set containing 75% and 25% of the samples, respectively.

2.1.2. Pretraining

The source speech is encoded at a frequency of 25 kHz for the pre-processing stage. The goal of this step is to reduce the amount of data in the feature map to speed up the learning process. In addition, STFT is used to convert a time domain signal to a spectral domain signal. At this stage, each file is divided into 10ms of Hamming window segments with 50% overlap between consecutive Windows. Finally, the spectrogram is changed to the same size of 60*155 points to get rid of the useless part that doesn't contain any information. In this case, useless noise is discarded and significant signs appear.

2.2. CNN architecture

CNN Is represented by an input layer and several hidden layers. Each individual layer consists of a convolutional layer H and a merging layer V. The input feature map is defined as $V_l(l = 1, ..., L)$, and the convolutional feature map is defined as $H_k(k = 1, ..., K)$. The filter weights are common to all units on the convolutional layer, calculated as,

$$h_m^k = \sigma(\sum_{l=1}^{I}\sum_{n=1}^{N_W} v_{l,n+m-1}w_{l,n}^k + w_0^k) \tag{1}$$

where $v_{l,m}$ element of the m-th unit of l-th input layer V, and h_m^k element of m-th block of the k-th convolutional layer H. N_w is defined as the size of the filters, $w_{l,n}^k$ is n-th unit of weight and $+w_0^k$ is the 0-th unit of weight.

In this procedure, objects are detected locally and automatically using shared weights across the feature map.

To reduce resolution in convolutional plys and reduce computational complexity, a union of convolutional maps is used. The maximization or averaging function is usually used to build the unifying layer. In this case, set G as the size of the merging window using the maximize function, and the element on the merging layer is defined as,

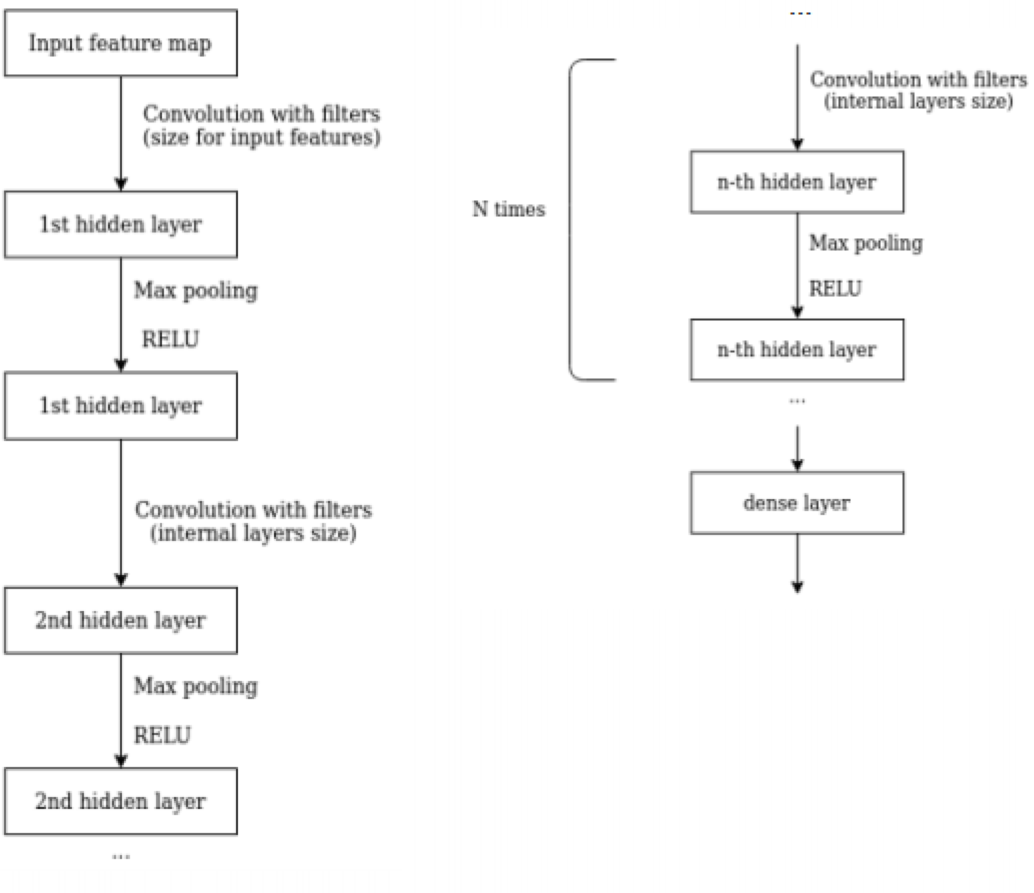

Figure 2: Network architecture.

$$p_m^k = max_{n=1}^G h_{l,(m-1)\times s+n} \tag{2}$$

where s is the step of the merging window moving in the convolutional layer and other variables are defined above.

The experimental network shown on figure 2 contains 10 hidden layers. In the first hidden layer, the filter size is 8*3, and the step is 1. The size of the merging window is 4*4 and step 1. After the first hidden layer, each layer was collapsed by 8 filters with the shape 8*3*8 and step 1. The size of the unifying windows is 4*4 and the RELU activation function for the entire neural network. Finally, the feature map is formed into a dense layer (a fully connected layer) to train the classification model. L2 regularization is used to solve the problem of retraining. Parameters such as pitch, size of filters in each layer, and the number of layers can be changed and should be selected depending on the signal features used. In this paper, we also studied networks with the configurations shown in table 1. The rectangular filter window is used because of the specific characteristics of the spectrograms.

Table 1
Configuration of the studied networks

Configuration	Input layer	Hidden layers
Proposed	Convolutional: 8*3*1 Pooling: 4*4*1	Convolutional: 8*3*8 Pooling: 4*4*1
Big filters	Convolutional: 16*6*1 Pooling: 8*8*1	Convolutional: 16*3*16 Pooling: 8*8*1
Small filters	Convolutional: 4*2*1 Pooling: 2*2*1	Convolutional: 16*3*16 Pooling: 2*2*1

2.3. Preprocessing

Deep learning is a "black box" that requires a large amount of data and processes to adjust the weight. In turn, Bayesian methods are reliable and interpretable on small amounts of data, which is exactly what deep learning methods lack.

To combine the complementary advantages of these two methods, generative models have been developed to improve the effectiveness of deep learning on small data sets and eliminate overfitting problems. In deep learning structures, a section of the weight space is detected by a generative model, which helps the network quickly converge to a global minimum. The convolutional restricted Boltzmann machine (CRBM) is a typical generative model and is an extension of RBM with visible and hidden layers as images that is suitable for CNN settings. The model is trained to reach a state of thermal equilibrium, which is the deepest energy minimum state. In this state, hidden layers can model the structure of input data.

The CRBM consists of two layers: the visible (input) layer V and the hidden (convolutional) layer H. Similar to the CNN setting, the weights W^k between the input layer and the convolutional layer are distributed among all elements in the hidden layer. Hidden elements are binary, while visible elements can be real or binary. Assume that the size of the visible layer is N_V, and the size of the hidden layer is N_H. There are K weights and each weight W_k is collapsed with the visible layer, and there is an offset b_k for each weight and an offset c for the visible layer. An energy function with a binary input is defined as,

$$E(v,h) = -\sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \tag{3}$$

An energy function with a real input is defined as

$$E(v,h) = \frac{1}{2}\sum_{i}^{N_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \tag{4}$$

Joint distribution is defined as,

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h)) \tag{5}$$

3. Results

Sensitivity shows the effectiveness of detecting abnormal voice files, and specificity shows the proportion of correctly detected healthy voice files. The accuracy (P) and F1-score (F1) are presented below, where the accuracy shows the proportion of the corresponding pathological voice files.

$$SN = \frac{TP}{TP + FN} \tag{6}$$

$$SF = \frac{TN}{FP + TN} \tag{7}$$

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$F1 = \frac{2p \cdot SN}{P + SN} \tag{9}$$

True negative (TN) means that healthy voice recordings are correctly identified. True positive (TP) means that abnormal voice recordings are correctly identified. False-negative (FN) indicates that abnormal voice recordings are detected incorrectly and false-positive (FP) indicates that voice recordings were detected incorrectly.

There is also a difference in the operation of the CT system with and without pre-training. When using a CDN to initialize weights, the CNN setup becomes more reliable, with similar performance for the custom data set and the test data set. This shows that the CDBN can avoid overfitting problems to some extent. However, the accuracy on the test dataset is less when using pre-trained CDBN weights.

Similarly, the CRBM is trained using Gibbs block sampling[10] as an extension of the Gibbs sampling in RBM to maximize the similarity of the distribution between the construction visible layer and the input visible layer, and in this case achieve an equilibrium state. The stacks from the CRBM make up the CDBN. After the first CRBM layer is trained, activations are sent to the input of subsequent layers and the weights are " frozen”, and the remaining layers are processed in the same way. Since the visible layer in the first layer works with real data, Gaussian visible units are used for the first CRBM layer. After pre-training the weights in each layer, reverse propagation is applied to fine-tune the weights for a better classification result. Testing results are shown in tables 2 and 3.

Table 2
Testing results depending on the network architecture

	SN	SP	P	F1
Proposed	0.73	0.69	0.72	0.71
Big filters	0.73	0.72	0.73	0.71
Small filters	0.73	0.69	0.71	0.71

Table 3
Testing results depending on the CDBN usage

	CNN	CNN + CDBN
validation	0.65	0.67
testing	0.78	0.72

References

[1] K. Verdolini and L. O. Ramig, "Occupational risks for voice problems," Logopedics Phoniatrics Vocology, vol. 26, no. 1, 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, vol. 1, 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, no. 1, 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, no. 8, pp. 1622-1634, 2014/08/03 2014.

[5] A. Al-nasheriet al., "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification," Journal of Voice, vol. 31, no. 1, pp. 113.e9-113.e18.

[6] G. Muhammadet 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.

[7] 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 Toledanoet al., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99-109.

[8] O. Abdel-Hamid, A. r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, 2014.

[9] G. Hintonet al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition:

The Shared Views of Four Research Groups," IEEE Signal Processing Magazine, vol.29, no. 6, pp. 82-97, 2012.

[10] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," presented at the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, 2009.

[11] Wu H. et al. A deep learning method for pathological voice detection using convolutional deep belief networks //Interspeech 2018. – 2018.