Separation of Several Illnesses Using Correlation Structures with Convolutional Neural Networks

Attila Zoltán Jenei, Gábor Kiss, Miklós Gábriel Tulics, Dávid Sztahó

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok krt. 2, 1117 Budapest, Hungary, e-mails: jenei@tmit.bme.hu, kiss.gabor@vik.bme.hu, tulics.miklos@vik.bme.hu, sztaho.david@vik.bme.hu

Abstract: There is already a lot of research in the literature on the binary separation of healthy people and people with some illnesses that affects speech. However, there are only a few examinations where more illnesses are recognized together. The examination of the latter is justified by the fact that a person may suffer from several illnesses at the same time to a certain extent. In the present study, multiclass classification of depression, Parkinson's disease, and general voice disorders (organic and functional dysphonia) was performed using speech samples. Foremost, several acoustic features were examined as input (such as Mel-Frequency Cepstral Coefficients (MFCCs), mel-band energy values, formants and their bandwidths). Using the inputs, auto- and cross-correlation structures were formed as image representations and fed to a convolutional neural network (CNN). Parameter optimization of the correlation structures and the CNN model was applied to achieve the highest accuracy. Moreover, the result of the tuned process was compared to the result of a baseline process. Finally, multiclass (5 and 4 classes) classification was performed with the best parameters. The prominent feature set was the MFCCs (55.9% accuracy, 52.2% macro F-score) for 5 class classification. 64.3% accuracy and 60.0% macro F1-score was obtained for 5 classes after parameter optimization. For classifying 4 classes (merging dysphonic ones together), 74.9% accuracy and 71.7% macro F1-score was achieved.

Keywords: depression; voice disorders; Parkinson's disease; speech; Convolutional Neural Network

1 Introduction

Several illnesses have effect on speech production making speech a very important biomarker. There is much research in the literature on recognizing a disease while comparing samples with healthy control. However, some illnesses can occur at the same time, for example, Parkinson's disease is often accompanied

by depression. In this paper, we demonstrate multiclass classification process separating illnesses such as depression, Parkinson's disease, and dysphonia, supplemented with healthy class.

Depression is one of the most common psychiatric illnesses, affecting more than 300 million people worldwide. Nearly 800,000 people commit suicide each year according to World Health Organization (WHO) due to depression [1]. Possible triggers of depression can be stressful or negative life events, physiological disorders, social problems [2]. Early detection of the disease is not always clear, as its symptoms vary widely from individual to individual [3]. The diagnostic process of depression is further complicated by the fact that the person can be completely isolated from society [4].

Parkinson's disease is a neurological degenerative disease that mainly occurs in the elderly. The source of this illness is the death of dopamine-producing cells in the brain. Typical symptoms are resting tremor, muscle rigidity, instability, bradykinesia. The disease also affects the vocal cords and the muscles of the face, thus appearing during speech production [5]. The importance of its early diagnosis is given by the fact that it is currently an incurable disease, the progression, and symptoms of which can only be alleviated [6].

Dysphonia (the auditory-perceptual symptoms of voice disorders) is a disease that occurs regardless of age and gender causing changes in speech quality. It is observed with an increased frequency in people who use their voice heavily, such as singers and teachers [7]. It directly affects the patient's quality of life, which can also bring about isolation from society, triggers depression, anxiety. It can also present as an accompanying symptom of tumours, which can be fatal if not properly diagnosed and treated [8]. Dysphonia is classified as either an organic or a functional dysphonia, where organic dysphonia results from some sort of physiological change in one of the subsystems of speech, while the latter refers to a voice problem in the absence of a physical condition.

In the conference article [9], the three disease classes – mentioned above – were included in addition to the healthy control class. Approximately 270 features were calculated per recording, including voice quality measures (e.g., jitter, shimmer), pitch and intensity related measures, spectral indicators (e.g., formant frequencies, MFCCs), prosodic features (e.g. Pairwise Variability Indices [PVI]), energy metrics (e.g. Soft Phonation Index [SPI]). Parkinson's disease, depression, and general voice disorders were classified with 10 fold cross-validation among the healthy class. As a result, the accuracy ranged from 71.7% to 86.6%. The former result was achieved by the k-nearest neighbours (k-NN) classifier, the latter was accomplished with support vector machine (SVM) with radial basis function. In addition, when feature selection was used, the accuracy of the SVM with radial basis function improved from 86.6% to 88%.

In a previous research, we have already examined the recognition of these three disease classes (depression, Parkinson's disease, and dysphonia) using auto and

cross-correlation structures from a limited set of acoustic-phonetic features [10]. The correlation structure was created following the work of Williamson et al., who have already successfully applied this solution in several researches [11-12]. The eigenvalues of the structures were used as input in the classification process created in RapidMiner Studio. k-NN and SVM algorithms were executed with 10 fold cross-validation. 78% accuracy was achieved using formants frequencies, MFCCs, mel-energy values, and fundamental frequency together.

Correlation structures have been already used for feature selection (Parkinson's disease, dysphonia) and recognition (dysphonia) by examining the sum of the upper triangular of the correlation matrix structure [13-14]. However, the correlation matrices as images for CNN have not been studied yet.

Numerous publications have been already published in the literature reporting binary classifications for the disease classes presented here. In these, high classification accuracy has been achieved (above 85%).

In a previous publication, the automatic separation of depression and healthy control was performed with 83%-86% accuracy with SVM. In this research audio recordings from 48 depressed subjects were used [15]. In a Chinese study, depression was detected with 82% accuracy using male speech samples with a regression procedure (for females the accuracy was 75%) [16].

In the case of Parkinson's disease, higher accuracy values (around 90%) can be found with both the sustained vowels and continuous speech in the literature [17-20]. However, small Parkinson's databases were usually used. According to the mPower research, 5.826 participants were tested by their sustained "a" sound, 86% accuracy was achieved [21].

Features like jitter, shimmer, MFCCs, and formant frequencies were the most commonly used acoustic features in recognizing dysphonia [22-23]. Both sustained vowels and continuous speech were examined. High accuracy (above 90%) also can be achieved to recognize dysphonia [24].

In this work, correlation structures were also created from certain features, but were used as an image representation and were fed into a CNN for classification. The application of correlation structures as an image on convolutional networks is novel, such a process has not been studied in these disease classes yet. Respectively, there is less research in the literature for examining these three illness groups simultaneously. However, such an investigation is justified by the fact that these three groups of illnesses may even be present simultaneously in a person's speech [25-27]. Furthermore, these illnesses are rarely suspected in the early stages. Such a device may help point out any of them by using a speech sample at the general practitioner.

In this study, a baseline CNN model was created first and a 5-class classification (depression, Parkinson’s disease, organic-, functional dysphonia, and healthy) was performed on it using several features. Secondly, parameter selection was done in

the correlation structure and the CNN model using a specific group of features. Finally, 4-class classification was executed with the tuned correlation structure and model.

The content of the article follows the next structure: In Section 2, the speech databases are presented. The process and methods are described in Section 3. In Section 4, the result of multiclass classification and parameter tuning is summarized. In Section 5, a conclusion is drawn from the research and the results.

2 Speech Databases

The database contained speech samples of the three illnesses (Parkinson's disease, depression, voice disorders) and healthy recordings as a control class. Prior to each recording, the patients (and the control subjects) signed an informed consent in which they agreed to use their voice recordings for research purposes.

Each subject read out loud "The North Wind and the Sun" in Hungarian language, a text that is often used in speech technology research. This resulted in about a one-minute-long recording for each subject. The database contained recordings in which the subjects did not have any other illnesses (other than Parkinson's disease, depression, voice disorders) that could have affected his or her speech. The presence of one disease (exclusion of other diseases) is certified by the doctor treating the patient. Audio materials were recorded at a sampling frequency of 44.1 kHz with a clip-on microphone in a quiet room. The recordings were stored in 16 bits in PCM format.

2.1 Depressed Speech Database (DE)

Several versions of BDI (Beck Depression Inventory) questionnaire were created. The latest version of which was published in 1996, named BDI-II was used in this research [28]. This version consists of 21 questions (0 to 3 score for each question). Speech recordings from people suffering from depression were approximately evenly distributed among the depression severity categories defined by the BDI-II (Beck Depression Inventory-II) as mild depression (score: 14-19), moderate depression (score: 20-28), and major depression (score: 29-63). Below the score of 14, the patient is considered healthy.

Speech samples from individuals suffering from depression were collected from the Psychiatric and Psychotherapy Clinic of Semmelweis University, Budapest.

A total of 91 speech samples were used from the Depressed Speech Database: 58 female subjects (mean BDI score: 27.6 (±9.3); mean age: 37.5 (±16.7) and 20 male subjects (mean BDI score: 26.6 (±8.6); mean age: 40.6 (±15.9)).

2.2 Voice Disorder Speech Database (UD)

Speech samples were recorded from patients diagnosed with different voice disorder by the Outpatients' Clinic of the Head and Neck Surgery Department of the National Institute of Oncology, Budapest.

The voice disorder database included patients' voice suffering from disorders such as functional dysphonia, recurrent paresis, tumours at the vocal tract, cysts, tract stenosis, vocal node, laryngitis, laryngeal paralysis, spasmodic dysphonia. Overall, these were divided into two major groups: organic dysphonia (OD) and functional dysphonia (FD). Together, OD and FD form the UD database.

The RBH (Roughness, Breathiness, Hoarseness) scale describes the severity of voice disorders that is widely used in Hungary [29]. The scale scores the roughness, breathiness, and hoarseness of the voice with integers between 0 and 3. The integer 3 is the most severe category. The severity of dysphonia was determined by the clinician who made the diagnosis during the consultations.

167 recordings (74 men and 93 women) were used from OD, while 68 (20 men, 48 women) were used from FD. Their mean ages were 51.6 (OD) and 55.8 (FD) years, respectively, and their standard deviations were 14.4 (OD) and 16.1 years (FD).

The hoarseness (H) value was used to describe the severity of the voice disorder. The mean hoarseness of functional dysphonia (FD) was 1.5 and the standard deviation was 0.7 for male subjects. For women, the mean was 1.3 and the standard deviation was 0.6. For organic dysphonia (OD), the mean of H was 2.1 and the standard deviation was 0.9 for male patients. For women, the values were 1.8 and 0.8, respectively.

2.3 Parkinson's Disease Speech database (PD)

Audio recordings of patients diagnosed with Parkinson's disease (PD) were collected from two locations in Budapest: Semmelweis University (25 recordings) and Virányos Clinic (55 recordings).

H&Y (Hoehn and Yahr) scale was used to describe the severity of the disease, which ranges from 1 to 5 [30]. The 5 indicates the most severe condition, while a 1 indicates mild symptoms. Furthermore, the scale is non-linear, from which it follows that H&Y 2 does not present twice as severe symptoms as H&Y 1.

80 speech samples were collected from patients with PD: 43 males (mean H&Y score: 2.7 (±1.2); mean age: 62.6 (±13.5)) and 37 females (mean H&Y score: 2.6 (±1.2); mean age: 65.2 (±9.2)).

2.4 Healthy Control Database (HC)

In addition, a database of healthy people's recordings was also created as a control group (HC). According to their own statement, healthy individuals did not have any illnesses (and have not been diagnosed with any known illnesses) that would affect their speech at the time of recording.

140 healthy speech samples were recorded: 85 female speakers (mean age: 49.6 (±15.2)) and 55 male speakers (mean age: 51.4 (±21.6)).

3 Methods

The examination process is illustrated in Figure 1. Firstly, acoustic features were obtained from the speech recordings. From these, auto and cross-correlation structures were generated. Finally, classification with the help of CNN was executed.

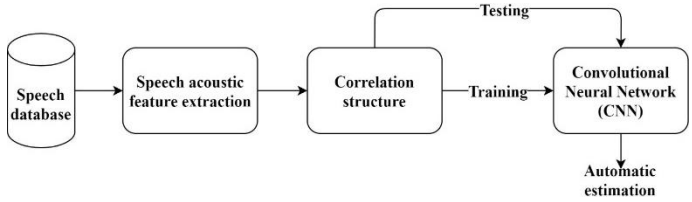

Figure 1
Outline of the applied process. (Speech database, feature extraction, correlation structure, training / testing on convolutional network)

The two-dimensional correlation matrices were input into a 2D convolutional neural network. After training the model, testing was done for automatic estimation.

Five feature sets were examined with a baseline process. From here, one set of features is selected for further studies. Furthermore, parameter optimization was performed on the correlation structure as well as on the machine learning model. In the tuned process, multiclass classification (with 5, 4 classes) was finally executed.

3.1 Acoustic Features Extraction

Before calculating the acoustic features, the speech samples were normalized to the peak. Then, acoustic features were calculated in a 50 ms Hamming window with the time step 10 ms using Praat software [31]. With this technique, a time series (later on referred as a vector) can be assigned to each feature per recording.

Then, the following speech acoustic features were obtained [32-33]:

Mel-Band Energy Values: The frequency range of the speech can be converted to a mel scale, from which mel-bands can be derived. The energy spectrum of speech can be passed through on these mel-bands, which resulted in cumulative energy values. The first 27 mel-band energy value was calculated from 100 Hz. Further on it is referred to as Melfilters.

Mel Frequency Cepstral Coefficient: This is determined from the power spectrum by summing the energy values within a defined mel-bands. Then, the discrete cosine transform of its logarithm value is calculated. The values of the first 14 coefficients were determined. These are hereinafter referred to as MFCCs.

Formant frequencies: The maximum amplitude's locations of the spectral envelope curves of the overtone beams amplified by human resonator cavities are called formant frequencies. The first three formant frequencies were calculated, which are hereinafter referred to as Formants.

Bandwidth of formant frequency: Bandwidth means the frequency range measured at a decrease of 3 dB from the amplitude peak of the formant frequency. The bandwidths of the 1^{st}, 2^{nd} and 3^{rd} formant frequencies were calculated, which are hereinafter referred to as Bandwidths.

Finally, formant frequencies and their bandwidths were also used in a combination as a fifth set of features, referred to as Form-Band. So that set included Formants and Bandwidths vectors.

Melfilters and MFCCs were calculated from the total speech sample, while formant frequencies and their bandwidths were calculated from the voiced sections. Thus, the vectors of MFCCs had the same length as Melfilters'. The length of the formant frequency vectors and the bandwidth vectors were also the same.

Where the feature extraction program could not determine a value that data itself was removed from the vector and also deleted from the other feature vectors (in the same set) on the same index even if it was a numeric value. Thus, the feature vectors did not shift relative to each other in time.

As a summary, Table 1 contains the extracted features, the name of the feature set, and the number of vectors in a set.

Table 1
Extracted acoustic features with the Praat software

Feature	Name of set	Number of vectors
Mel-Band Energy Values	Melfilters	27
Mel Frequency Cepstral Coefficient	MFCCs	14
Formant frequencies	Formants	3
Bandwidth of formant frequency	Bandwidths	3
Combination of formant frequencies and their bandwidths	Form-Band	6

3.2. The Structure of Correlation Matrices

Instead of using one single vector, several new vectors were created by shifts along time. At each shift, the elements were displaced by a certain extent (hereinafter referred as displacement rate) so that the last elements were placed at the beginning of the vector. At each shift, a new vector is produced. A general approach is shown in Eq. (1), where X_0 is the original feature vector, $x_1, x_2, \dots, x_m$ are its vector components (features from time to time). X_1 is a new vector with one element (displacement rate is 1) shift. X_i is the i^{th} new vector after i element (displacement rate is 1) shift.

$$\begin{aligned} X_0 &= [x_1, x_2, \dots, x_m] \\ X_1 &= [x_m, x_1, \dots, x_{m-1}] \\ X_i &= [x_{m-(i-1)}, x_{m-(i-2)}, \dots, x_m, x_1, \dots, x_{m-i}] \end{aligned} \tag{1}$$

Pearson's correlation coefficient was used to describe the linear relationship of two feature vectors [34]. Calculating this correlation coefficient between the two original and their shifted feature vectors, a matrix can be filled.

Denoting (k - 1) as the number of shifts, a submatrix of size $k \times k$ can be created using two feature vectors and their shifted variants from one set. This is shown on the right side of Figure 2. The rows stand for the first while the columns stand for the second feature vector. The first row and column indicate the original vectors whilst the other rows and columns represent the shifted vectors. The cells of the matrix contain the correlation coefficients of the two specific feature vectors. For example, the cell of the 3rd row and 2nd column (marked by a blue rectangle in Figure 2) includes the correlation coefficient of the two times-shifted Vector 2 and 1 time shifted Vector 1, respectively. For instance, Vector 1 can be the first and Vector 2 can be the second formant frequency vector.

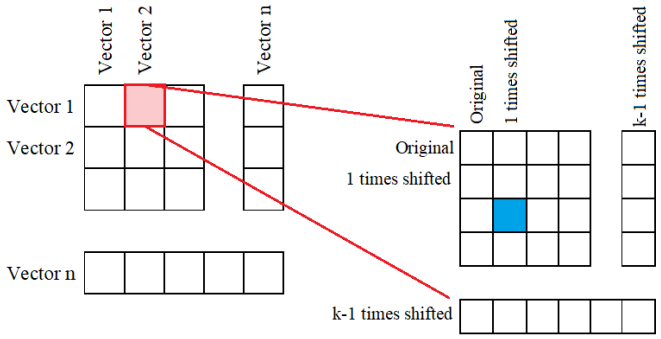

Figure 2

Structure of the correlation matrix: on the right side the submatrix of two feature vectors (size: $k \times k$). On the left there is the complete structure using a feature set with n vectors (size:$(k \times n)(k \times n)$).

One set of features was used up at once to create a correlation structure. Denoting the number of vectors in a set with n, the total size of the correlation structure is $(k \times n)(k \times n)$.

The total correlation structure is shown on the left side of Figure 2. The constructed structure is symmetrical, with autocorrelation coefficients in the main diagonals and cross-correlation coefficients in the sub-diagonals.

9 times shift (k = 10) with displacement rate 1 were set to create a baseline process. These baseline parameters (for the correlation structures and CNN) were successfully applied in preliminary research [35]. Later on, 4 times (k = 5) and 14 times shift (k = 15) were also examined with the displacement rate 1, 4, and 8 as parameter tuning.

One correlation structure was constructed for each person from each feature set (Overall, 5 correlation structures were available for each person). These as image representations were the input to the classification algorithm.

3.3 Construction of CNN Model

A simple CNN was created in Python (version 3.7.0) using Tensorflow (version 1.12.0). The baseline parameters are based on preliminary research [35].

A sequential CNN model was created with two convolutional layers, followed by a maxpooling, a flatten and a dense layer. The arrangement of the CNN layers is shown in Figure 3.

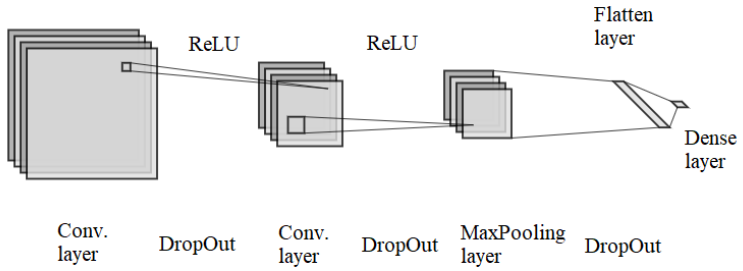

Figure 3
The structure of the CNN model: two Convolutional layers, one MaxPooling, Flatten and Dense layer

Correlation structures with the size of $(k \times n)(k \times n)$ described above were used as input to the CNN. 32 kernels were used in the first Convolutional layer. Kernel size $k \times k$ and stride k have been set according to the size of the submatrices of the input images.

The kernel size and stride of the second Convolution layer were chosen so that 2×2 size matrices were at the output of the layer. The size of the stride was equal to any dimension of the square kernel. 32 kernels were also used here.

The pool size of the MaxPooling layer was 2 × 2 as default.

ReLU (Rectified Linear Unit) activation functions were set up after the first two Convolutional layers and 25% DropOut regulations after each of the first 3 layers.

Finally, the output values of the Dense layer were converted into probability values with the SoftMax function. A vector with x components resulted for each test subject, where x denotes the number of classification categories. The predicted class is the one having the highest probability output.

ADAM optimization was applied during training sessions. Herewith, the automatic adjustment of the learning rate is realized taking into account the cost function. The cost function used here was the categorical cross-entropy [36].

For pre-processing, the imported data was shuffled wherein one correlation structure belonged to one subject. Normalization between 0 and 1 was also done on the elements of the matrices.

3.4 Tests and Evaluation Methods

With the created process (feature extraction from speech recordings, the built-up of the correlation structures, creating the CNN model), the following examinations were performed. The first two tests were executed with 5 classes: DE, PD, FD, OD, HC. Then the last test was performed with the database where FD and OD were merged to UD.

Leave-one-out cross-validation (LOOCV) was used for model evaluation for all tests. During this process, one subject is selected as the test element while the remaining samples are used as the training set. This is repeated until every sample was a test element. This means the training and testing process is repeated as many times as many samples are in the database. Moreover, the testing and training set were always disjoint.

For evaluation, confusion tables were created from the output of the CNN models. Metrics such as recall, precision, accuracy, and F1-score were derived.

a) Examination of feature sets: the 5 feature sets were tested separately in the baseline model. This gave sequential results on which features most appropriate for separation using this certain process.

b) Parameter optimization: In this test, the baseline process was tuned. Specifically, the number of shifts and the displacement rate were changed in the correlation structure. By changing the displacement rate, the parameters of the neural network model were not changed. However, by changing the number of shifts, the kernel size and strides of the first convolution layer were adjusted as shown in Table 2. Finally, 4 time (k = 5), 9 time (k = 10) and 14 time (k = 15) shifts were examined. The displacement rate was set as 1, 4, 8. The number of kernels remained 32 for both Convolutional layers in this case.

Table 2
The kernel size and stride of the first Convolutional layer based on the number of shifts

Number of shift	k	Kernel size	Kernel stride
4	5	5 × 5	5
9	10	10 × 10	10
14	15	15 × 15	15

After choosing the right parameters for the correlation structures, the CNN parameter settings followed. The number of iterations during training and the number of kernels were changed in the new CNN model. The number of iterations was set to 25, 50, 75, 100, 125, 150, and the number of kernels was set to 16, 32, 64, 128. The kernel numbers were chosen so that the first convolution layer had half the kernel number as had the second convolution layer. Thus, kernel numbers 16/32, 32/64, and 64/128 were used, where the first number is the kernel number of the first convolution layer and the second is the kernel number of the second convolution layer.

The parameter optimization was done before the training cycle and then it was tested by all the subjects separately. With this in mind, the separation of a third independent set was not necessary as the test set was already independent for the models.

c) 4 classes classification: By combining organic (OD) and functional (FD) dysphonia, the general voice disorders (UD) group was created to investigate the 4-class classification in the optimized process.

4 Results

4.1 Examination of the 5 Feature Sets

The 5 feature sets were examined in the baseline process (9 times shift, 1 displacement rate, 32 kernels, 100 iterations) based on accuracy and F1-score. All subject was used from the 5 classes. Results are shown in Figure 4.

Accuracy values ranged from 43% to 56% for all feature sets, while macro F1-score values ranged from 36% to 52%. The highest accuracy and macro F1-value were achieved with the MFCCs feature set (55.9% accuracy, 52.2% macro F1-value). Melfilters achieved the second-best accuracy (51.1%) and macro F1-score (47.4%).

Using the formants and their bandwidths separately, we obtained an accuracy of 43.0% (formants) and 45.5% (bandwidths). While using them together, the results improved (52.0% accuracy, 44.3% macro F1-score).

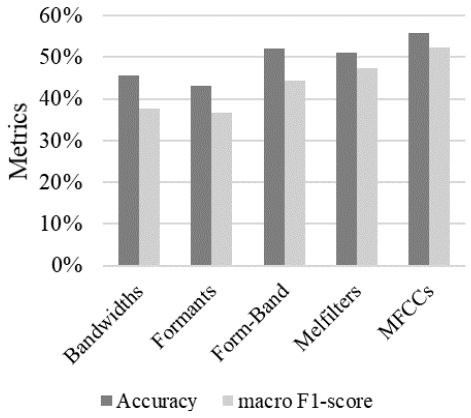

Figure 4
Achieved accuracy and macro F1-score with different feature sets on the baseline process using 5 classes

Table 3 shows the confusion matrix created from the feature set that achieved the highest accuracy (MFCCs). The columns are the original classes and the rows are the classifier's decisions. Precision and recall by classes are also noted.

The recall of the DE class was low (38.5% recall), while the precision was high (64.8% precision) compared to the other classes. The recall of the HC class was 73.6%. On the other hand, many samples from originally positive classes were classified as healthy, resulting in a 53.1% precision.

It can also be seen that subjects with functional dysphonia tended to be classified as organic dysphonia or healthy (10.3% recall, 50% precision).

Table 3
The confusion matrix derived from the MFCCs feature set. The columns represent the original classes the rows represent the decision of the algorithm.

		Original classes					
		HC	DE	FD	OD	PD	Precision
Predicted classes	HC	103	26	26	30	9	53.1%
	DE	8	35	1	3	7	64.8%
	FD	2	0	7	5	0	50.0%
	OD	18	18	30	118	22	57.3%
	PD	9	12	4	11	42	53.8%
	Recall	73.6%	38.5%	10.3%	70.3%	52.5%	

4.2 Parameter Optimization

MFCCs were selected to adjust the correlation structure and parameters of CNN to achieve a better separation of classes.

The results obtained by changing the number of shifts (k) and displacement rate are shown in Figure 5. The accuracy is given on the left diagram, the macro F1-score is given on the right diagram. The displacement rates are on the category axis while the shades of the bars indicate the number of shifts.

According to Figure 5, it is worthwhile to use a correlation structure with a higher displacement rate. However, further changes were not experienced with the displacement rate of 8. Changing the number of shifts is significant at a low displacement rate. While at a higher displacement rate, changing the number of shifts will only cause small changes in the metric values.

The highest, 61.7% accuracy was achieved at the displacement rate 4 with 15 shifts. Using macro F1-score, 55.5% is reached as the peak at the displacement rate 8 with 10 shifts.

The displacement rate 8 and 4 times shift were selected for further examination because at this displacement rate, all three shift numbers gave similar results. Nonetheless, the CNN parameters increased polynomial by linearly increasing the number of shifts.

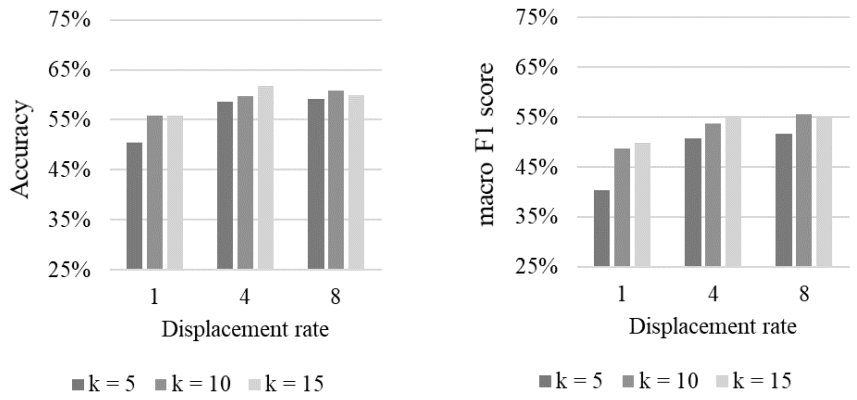

Figure 5

Results obtained by varying the number of shifts and displacement rates in the correlation structure. The left side chart shows the accuracy, while the right side chart shows the macro F1-score.

In the case of the CNN model, several parameters can be set, from which the number of iterations during the training and the number of kernels of two convolutional layers were selected for analysis.

The results obtained using different iteration numbers are shown in Figure 6. The horizontal axis shows the number of iterations, the vertical axis the percentage of accuracy.

The mean (black line) and standard deviation (grey band) of the accuracy was calculated and plotted during the training process. The accuracy of the test set was also plotted (grey curve). In the latter case, the standard deviation could not be calculated.

Based on Figure 6, the accuracy of both the test set and the training set increases. Over 125 epochs, a decrease can be observed at the test set accuracy. Thus, the iteration number in the CNN model was set from 100 to 125 in the following experiments.

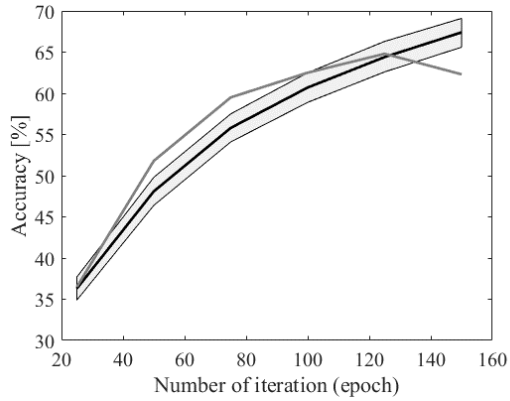

Figure 6

The accuracy of the training (black line with grey band) and test set (grey curve) as a function of the epoch number for 5 classes

The second parameter was the kernel number of the two convolutional layers to set. Kernel numbers were determined as the power of two so that the kernel number of the second convolutional layer was twice that of the first. Based on this, 16/32, 32/64, and 64/128 kernels were applied with 125 iterations (k = 5, displacement rate 8, MFCCs feature set).

The results are shown in Figure 7, where the kernel numbers are shown on the category axis. The vertical axis shows the percentages of the accuracy and F1-score.

The value of accuracy ranged from 61.2% to 64.3% in this examination. The maximum of 64.3% was reached by setting the 32/64 kernels. Similarly, the macro F1-score had a maximum of 60.0% at 32/64 kernels. The lowest value of the macro F1-score was 54.6% at 16/32 kernels. The precision averaged along the classes varied over a narrow range, from 60.0% to 61.8% along the category axis. Recall increased from 55.2% to 60.0% at 32/64 kernels.

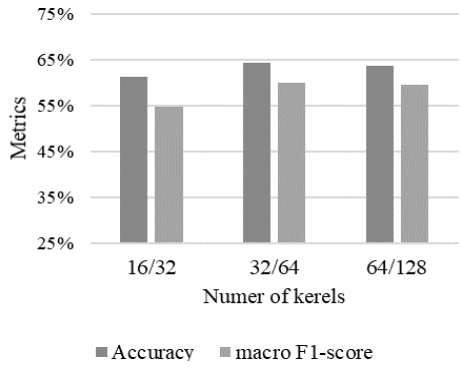

Figure 7

The classification results by choosing different kernel numbers. The first number is the kernel number of the first Convolution layer, the second is the kernel number of the second Convolution layer on the category axis.

Summarizing the results obtained by setting the two parameters of the network: the maximum of 125 iterations are worth using in the present construction. Setting 32/64 kernels brought the highest accuracy and macro F1-score in the present method. It should also be noted that this choice of kernel numbers increased the number of free parameters in the model approximately polynomially. Thus, setting these parameters can also be considered when designing such an experiment.

4.3 4 Classes Classification

Based on the results of the classification with the baseline process, organic and functional dysphonia were difficult to distinguish from each other. Therefore, these two groups were combined and examined as the general voice disorder group. This test has been done by applying the optimized parameters (iteration: 125, 32/64 kernels, k = 5, displacement rate 8, MFCCs feature set) on the system.

Recognition of depression was reduced by 7 samples, healthy by 4 samples, and Parkinson's disease by 1 sample in the classification of 4 classes compared to 5 classes. However, the recognition of UD was improved by 70 samples. The overall results can be seen in Table 4. Accuracy 64.3% was achieved for 5 classes and 74.9% for 4 classes on the tuned system. Macro F1-score increased from 60.0% to 71.7% by using 4 classes instead of 5 on the optimized process.

Table 4

Result of 5 and 4 classes classification with the optimized process using MFCCs

	Accuracy	**macro F1-score**
5 classes	64.3%	60.0%
4 classes	74.9%	71.7%

Discussion

Using the baseline process, the MFCCs feature set performed the best (55.9% accuracy, 52.2% macro F1-score). The MelFilters feature set resulted in the second-best output (47.4% accuracy, 51.1% macro F1 value). Its drop compared to the MFCCs is probably due to the fact that the 27 mel-band energy values contain everything up to 8 kHz, including signals that do not contribute to separation but are interfering.

Furthermore, the Form-Band features indicated that the combination of formant and bandwidth could improve the separation of the classification algorithm compared to applying them separately.

Increasing the number of shifts increased accuracy and macro F1-score. A possible reason for this may be that the first convolution layer may have performed better convolution from multiple samples (from a larger input context) than from a few samples.

Metrics also improved by increasing the displacement rate, but a slight decrease was already experienced at a displacement rate of 8. The decrease may be due to the disappearance of the correlation relationship between the two feature vectors in the sub-diagonals. Thus, strong correlations in the structure are limited to the main diagonal, which adversely affects the classification.

The highest test accuracy was obtained at 125 iterations, where even the results of the training and test set together progressed within the deviation band. The risk of overfitting on the training set increases at higher iterations. Also, the result of the test set has already decreased. At a low number of iterations, the accuracy of the test set changes more dynamically compared to the training set. This can presumably be caused by underfitting.

Changing the kernel numbers brought a bigger change in the macro F1-score compared to the accuracy value. Their highest performance (accuracy: 64.3%, macro F1-score: 60.0%) occurred at 32/64 kernel number. Using a higher kernel number (than 32/64) did not improve the classification, but it did increase significantly the running time of the training.

Combining OD and FD together as UD improved their correct recognition (accuracy: 74.9%, macro F1-score: 71.7%) while maintaining the optimized parameters. This is certainly influenced by the relatively large number of elements in the UD class. In contrast, the average improvement across classes is greater for 4 classes than 5 classes.

Conclusions

In the present work, the recognition of depression, Parkinson's disease, and general voice disorders were examined using a method in a new approach. In this procedure, acoustic features were calculated from speech. Component shifts were performed on the feature vectors, from which a correlation matrix was created.

These matrices were the input of a CNN model to execute the separation. First, a baseline process served the feature selection purpose from several feature sets. Secondly, parameter optimization of the correlation structures and the CNN model was also performed. Finally, 4 and 5 class classification were performed.

The advantage of this method is that it does not require more complex speech processing (such as segmentation). Furthermore, the convolutional network itself extracts the essential information from the image representations.

Also, the classification results of 4 classes can be compared to the results discussed in [10]. Higher accuracy (74.9%) was achieved here against 69.4% in [10] using only MFCCs. Unfortunately, exceeding 86.6% that had been achieved in the [9] was not successful. However, multiple features were applied there while only MFCCs were applied in the present study. Moreover, many features required segmentation in [9] while the presented method here does not need segmentation which can be a huge advantage.

Acknowledgement

Project no. K128568 has been implemented with the support provided by the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme. The research was partly funded by the CELSA (CELSA/18/027) project titled: “Models of Pathological Speech for Diagnosis and Speech Recognition”.

References

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators: Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017, The Lancet, 392(10159), pp. 1789-1858, 2018, https://doi.org/10.1016/S0140-6736(18)32279-7

[2] J. W. Kanter, A. M. Busch, C. E. Weeks, S. J. Landes: The nature of clinical depression: Symptoms, syndromes, and behavior analysis. The Behavior Analyst, 31(1), pp. 1-21, 2008, https://doi.org/10.1007/bf03392158

[3] G. S. Malhi, J. J. Mann: Depression, The Lancet, 392(10161), pp. 2299-2312, 2018, https://doi.org/10.1016/S0140-6736(18)31948-2

[4] L. Ge, C. W. Yap, R. Ong, B. H. Heng: Social isolation, loneliness and their relationships with depressive symptoms: A population-based study, PLoS ONE, 12(8), e0182145, 2017, https://doi.org/10.1371/journal.pone.0182145

[5] E. A. C. Pereira, T. Z. Aziz: Parkinson’s disease and primate research: Past, present, and future, Postgraduate Medical Journal, 82(967), pp. 293-299, 2006, https://doi.org/10.1136/pgmj.2005.041194

[6] L. V. Kalia, A. E. Lang: Parkinson's disease, The Lancet, 386(9996), pp. 896-912, 2015, https://doi.org/10.1016/S0140-6736(14)61393-3

[7] A. C. Gama, J. N. Santos, E. F. Pedra, A. T. Rabelo, M. C. Magalhães, E. B. Casas: Vocal dose in teachers: correlation with dysphonia, CoDAS, 28(2), pp. 190-192, 2016, https://doi.org/10.1590/2317-1782/20162015156

[8] R. J. Stachler, D. O. Francis, S. R. Schwartz, C. C. Damask and et al.: Clinical Practice Guideline: Hoarseness (Dysphonia) (Update), Otolaryngology-Head and Neck Surgery, 159(1), pp. 1-42, 2018, https://doi.org/10.1177/0194599817751030

[9] D. Sztahó, G. Kiss, M. G. Tulics, B. Hajduska-Dér, K. Vicsi: Automatic discrimination of several types of speech pathologies, in 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, pp. 1-6, 2019, https://doi.org/10.1109/SPED.2019.8906556

[10] D. Sztahó, G. Kiss, M. G. Tulics, K. Vicsi: Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features, in 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, pp. 1-4, 2018, https://doi.org/10.1109/TSP.2018.8441395

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, D. D. Mehta: Vocal and facial biomarkers of depression based on motor incoordination and timing, in Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 65-72, 2014, https://doi.org/10.1145/2661806.2661809

[12] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, D. D. Mehta: Vocal biomarkers of depression based on motor incoordination, in Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 41-48, 2013, https://doi.org/10.1145/2512530.2512531

[13] T. Bocklet, E. Noth, G. Stemmer, H. Ruzickova, J. Rusz: Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis, in 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, pp. 478-483, 2011, https://doi.org/10.1109/ASRU.2011.6163978

[14] T. Dubuisson, T. Dutoit, B. Gosselin, M. Remacle: On the use of the correlation between acoustic descriptors for the normal/Pathological voices discrimination, EURASIP Journal on Advances Signal Processing, 173967 (2009), 2009, https://doi.org/10.1155/2009/173967

[15] G. Kiss, K. Vicsi: Comparison of read and spontaneous speech in case of automatic detection of depression, in 8th IEEE International Conference on

Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, pp. 213-218, 2017, https://doi.org/10.1109/CogInfoCom.2017.8268245

[16] H. Jiang, B. Hu, Z. Liu, G. Wang, L. C. Zhang, X. Li, H. Kang: Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features, Computational and Mathematical Methods in Medicine, 2018, https://doi.org/10.1155/2018/6508319

[17] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig: Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease, IEEE Transactions on Biomedical Engineering, 59(5), pp. 1264-1271, 2012, https://doi.org/10.1109/TBME.2012.2183367

[18] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene: Detecting Parkinson's disease from sustained phonation and speech signals, PLoS ONE, 12(10), pp. 1-16, 2017, https://doi.org/10.1371/journal.pone.0185613

[19] A. Benba, A. Jilbab, A. Hammouch, S. Sandabad: Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease, in 2015 International Conference on Electrical and Information Technologies (ICEIT), Marrakeck, Morocco, pp. 300-304, 2015, https://doi.org/10.1109/EITech.2015.7163000

[20] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, S. Sapir: Early diagnosis of Parkinson's disease via machine learning on speech data, in 2012 IEEE 27th Convention Electrical and Electronics Engineers in Israel, Eilat, Israel, pp. 1-4, 2012, https://doi.org/10.1109/EEEI.2012.6377065

[21] B. M. Bot, C. Suver, E. C. Neto, et al: The mPower study, Parkinson disease mobile data collected using ResearchKit", Scientific Data, 3, 2016, https://doi.org/10.1038/sdata.2016.11

[22] J. P. Teixeira, P. O. Fernandes: Acoustic Analysis of Vocal Dysphonia, Procedia Computer Science, 64, pp. 466-473, 2015, https://doi.org/10.1016/j.procs.2015.08.544

[23] H. T. Lathadevi, S. P. Guggarigoudar: Objective acoustic analysis and comparison of normal and abnormal voices, Journal Clinical and Diagnostic Research, 12(12), pp. 1-4, 2018, https://doi.org/10.7860/JCDR/2018/36782.12310

[24] J. I. Godino-Llorente, P. Gomez-Vilda: Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, in IEEE Transactions on Biomedical Engineering, 51(2), pp. 380-384, 2004, https://doi.org/10.1109/TBME.2003.820386

[25] L. Marsh: Depression and Parkinson's disease: current knowledge, Current Neurology and Neuroscience Reports, 13, 2013, https://doi.org/10.1007/s11910-013-0409-5

[26] A. Hu, A. Hillel, W. Zhao, T. Meyer: Anxiety and depression in spasmodic dysphonia patients, World Journal of Otorhinolaryngology - Head and Neck Surgery, 4(2), pp. 110-116, 2018, https://doi.org/10.1016/j.wjorl.2018.04.004

[27] L. J. White, E. R. Hapner, A. M. Klein, et al.: Coprevalence of anxiety and depression with spasmodic dysphonia: a case-control study, Journal of Voice, 26(5), pp. 1-6, 2012, https://doi.org/10.1016/j.jvoice.2011.08.011

[28] A. T. Beck, R. A. Steer, R. Ball, W. F. Ranieri: Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients, Journal of Personality Assessment, 67(3), pp. 588-597, 2010, https://doi.org/10.1207/s15327752jpa6703_13

[29] T. Haderlein, C. Schwemmle, M. Döllinger, et al: Automatic Evaluation of Voice Quality Using Text-Based Laryngograph Measurements and Prosodic Analysis, Computational and Mathematical Methods in Medicine, 2015, pp. 1-11, 2015, https://doi.org/10.1155/2015/316325

[30] J. M. Rabey, A. D. Korczyn: The Hoehn and Yahr Rating Scale for Parkinson's Disease, in Instrumental Methods and Scoring in Extrapyramidal Disorders, Heidelberg: Springer, Berlin, Heidelberg, pp. 7-17, 1995, https://doi.org/10.1007/978-3-642-78914-4_2

[31] P. Boersma: Praat, a system for doing phonetics by computer, Glot International, 5(9/10), 341-345, 2002

[32] B. H. Story: Vowel and consonant contributions to vocal tract shape, The Journal of the Acoustical Society of America, 126(2), pp. 825-836, 2009, https://doi.org/10.1121/1.3158816

[33] J. Saini, R. Mehra: Power Spectral Density Analysis of Speech Signal using Window Techniques, International Journal of Computer Applications, 131(14), pp. 33-36, 2015, https://doi.org/10.5120/ijca2015907549

[34] M. M. Mukaka: Statistics corner: A guide to appropriate use of correlation coefficient in medical research, Malawi Medical Journal, 24(3), pp. 69-71, 2012, PMCID: PMC3576830

[35] A. Z. Jenei, G. Kiss: Possibilities of Recognizing Depression with Convolutional Networks Applied in Correlation Structure, In: 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), Milan, Italy, pp. 101-104, 2020, https://doi.org/10.1109/TSP49548.2020.9163547

[36] K. P. Murphy: Probability, in Machine Learning: A Probabilistic Perspective, (ed.) The MIT Press, Cambridge, Massachusetts, 2012