Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 164 (2019) 662–669

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019

Transfer Learning with AudioSet to Voice Pathologies Identification in Continuous Speech

Victor Guedes[a,b], Felipe Teixeira[a], Alessa Oliveira[a,c], Joana Fernandes[a], Leticia Silva[a,c], Arnaldo Junior[b], João Paulo Teixeira[a,d] *

[a] *Instituto Politécnico de Bragança, Bragança (IPB) 5300, Portugal*
[b] *Universidade Tecnológica Federal do Paraná, Câmpus Medianeira, Brasil*
[c] *Universidade Tecnológica Federal do Paraná, Câmpus Cornélio Procópio, Brasil*
[d] *Research Centre in Digitalization and Intelligent Robotics (CEDRI), Applied Management Research Unit (UNIAG), IPB, Bragança, Portugal*

Abstract

The classification of pathological diseases with the implementation of concepts of Deep Learning has been increasing considerably in recent times. Among the works developed there are good results for the classification in sustained speech with vowels, but few related works for the classification in continuous speech. This work uses the German Saarbrücken Voice Database with the phrase "*Guten Morgen, wie geht es Ihnen?"* to classify four classes: dysphonia, laryngitis, paralysis of vocal cords and healthy voices. Transfer learning concepts were used with the AudioSet database. Two models were developed based on Long-Short-Term-Memory and Convolutional Network for classification of extracted embeddings and comparison of the best results, using cross-validation. The final results allowed to obtaining 40% of f1-score for the four classes, 66% f1-score for Dysphonia x Healthy, 67% for Laryngitis x healthy and 80% for Paralysis x Healthy.

* Corresponding author. Tel.: +351 273 30 3129; fax: +351 273 30 3051.
E-mail address: joaopt@ipb.pt

1877-0509
10.1016/j.procs.2019.12.233

Keywords: Long Short Term Memory; Convolutional Neural Network; SVD; Deep Learning; Voice Pathologies Diagnose.

1. Introduction

Voice pathologies affect the speech performance of the patient, which causes disorders such as irritation, tiredness in speech, difficulties in word pronunciation, among other cases. Examples of related diseases, used in this work, there are three in particular: dysphonia, chronic laryngitis and paralysis of the vocal chords. Dysphonia is related to the difficulty in maintaining the voice and mainly associated with hoarseness. Chronic laryngitis also has symptoms of hoarseness, caused by long period inflammation of the larynx. Paralysis of the vocal cords mainly affects the speech and breathing of the patient, and, if both cords are paralyzed, the patient also presents loss of voice power.

The medical examination for the identification of these speech pathologies is somewhat invasive for the patient, and it is for this reason that the study and development of approaches with machine learning and deep learning has become more present during the last years.

Among the works developed, most of them have been evolved for the classification of innumerable pathologies using audios with sustained speech, mainly with the vowel /a/ [1] [2] [3], or using also other vowels such /i/ and /u/ [4] [5] [6] and getting good results. However, despite the scientific interests for the issue there is still very few the continuous speech databases with medical diagnose annotations available to research. Concerning the research of voice diagnose using continuous speech there is a few scientific research [7], despite the obvious higher practical interest. In this sense, this work aims to present a methodology for the classification of pathologies in continuous speech using transfer learning concepts for a database of audios without relations with voice diseases.

Nomenclature

LSTM	Long Short Term Memory
CNN	Convolutional Neural Networks
STFT	Short-time Fourier transform
PCA	Principal Component Analysis

2. Development

The idea to do a transfer learning is when there is a model already trained and that has a good performance in the problem to be solved [8]. In this sense, as this work is using phrases to detect voice pathology, it would be interesting to have a model trained in several pathological classes and that has a relation to voice, however this problem has not yet been solved. It was necessary to adapt the approaches that will be used.

The hypothesis of this experiment is based on if a pre-trained model can extract related characteristics that help in the response of a new problem. Thus, the pre-trained model used was VGGnet (named VGGish) with Google's AudioSet [9] [10].

AudioSet is a database with 2.1 million cataloged audios, which is equivalent to 5.8 thousand hours of audio divided into 527 classes. In these classes, there are present audios of music, voice, vehicles, musical instruments, among others [9].

Initially trained for image classification in the ImageNet challenge, the VGGnet model was adapted and trained in the AudioSet database. The VGGnet is composed of four convolutional layers and max polling, with ReLu activation functions, followed by two layers fully connected with ReLu and the embedding layer, which is the layer used to extract characteristic matrices. The VGGish template can be found at the github address available at https://github.com/tensorflow/models/tree/master/research/audioset.

There are some techniques for applying transfer learning and they can be applied in this model. Retraining the network for new classes by adding more fully-connected layers to the top of the model is an option. It is also possible to use the models as a characteristic extractor and then use these extractions as input from a new classifier to determine the class desire. The second approach was chosen.

In relation to the pathological basis, the German database Saarbrücken Voice Database (SVD) [11] with the sentence "*Guten Morgen, wie geht es Ihnen?"* ("Good Morning, how are you?") was used with a total of 70 dysphonia patients, 82 of chronic laryngitis, 197 of vocal cords paralysis and 632 for healthy subjects.

2.1. Preprocessing

In order to extract the characteristics using the VGGish model, a preprocessing was necessary. Initially, silence was removed from the beginning and end of each audio to prevent it from influencing the learning of the model. For this, it was first necessary to establish a decibel limit to be considered as silence, this number was defined by calculating the average of all the first decibels of the audios. Next, a method has been developed that traverses audio in 10 millisecond and analyzes whether the decibel value of this chunk is smaller than -41dB (silence limit) set, if yes, recognizes silence. This is done until it finds a higher value than the one established (example -20dB).

The same process is performed to silence at the end of the audio (using -41dB), passing the inverted audio vector. Silences in the middle of the signal were not analyzed or removed, as this is part of speech.

The spectrograms are then extracted using the standard preprocessing required for use of VGGish [10]. First, all audios are converted to 16 kHz sampling frequency, and from stereo to mono, the spectrogram is calculated using Short-Time Fourier Transform (STFT) [12], using a window size of 25ms and this being traversed by 10ms windowed with a Hanning window. Subsequent to this a mel spectrogram is calculated with 64 mel bins for the range of 125 to 7500 Hz, which is also made a mel-spectrum (mel-spectrum + 0.01) with offset to avoid zero logarithm. Finally, these spectrograms are defined in non-overlapping samples of 10 milliseconds, different from the 0.96 standard, because the size of the German base audios are smaller. Each example is defined in 64 Mel bands for 96 frames of 10ms each. This will be the input of the VGGish network [10].

2.2. Post processing

The standard AudioSet implementation extract the embedding from VGGish and after apply a Principal Component Analysis (PCA) transformation [13] for the AudioSet PCA matrix and then use quantization [14] with a range of 0 to 255 transforming the data to 8 bits so that the database is compatible with the database of YouTube-8M [10]. However was not necessary to do the same process in the German database, because it does not need to be compatible with the YouTube-8M database. Thus, was used raw embeddings from VGGish, which generated matrix in the format (N, 128) where N is the number of examples extracted from the spectrograms. This value of N varies according to the size of the audio, where the highest audio has a value of 363. In this sense, several tests with various sizes of N have been made, and the best of them is set at 80. Thus, the format of input of the following networks are 80x128.

With the data extracted, it is now possible to make the classification within the four pathologies. For this, two models of Deep learning networks were developed that support data analysis with temporal variability. The first is the Long Short Term Memory (LSTM) based on recurrent networks [15] [16]. This is composed of an input layer of Batch Normalization [17] with 80x128 format, followed by two LSTM layers with 64 and 32 neurons, respectively, and hyperbolic tangent activation function, and a fully connected layer with 4 neurons with Softmax function representing the output.

The second network is based on one-dimensional Convolutional Neural Network (CNN) [18]. It is implemented by a Batch Normalization entry layer equal to LSTM, followed by a 1D convolution layer with 32 neurons and kernel size 3, and hyperbolic tangent activation function. It is followed by a layer of MaxPooling1D with size 2, another convolution 1D with the same previous configurations, followed by a layer of GlobalMaxPooling 1D, to transform the data of three dimensions for 2 dimensions. IT fallows a layer of Dropout [19] with 0.2 value to avoid overfitting and finally the output layer equal to the layer used in the LSTM.

The two networks were trained in 1000 epochs, with categorical cross entropy error function [19], with Bach size of size 32. In addition, the number of instances was decreased in healthy and paralysis of the vocal cords so that the

training data were approximately balanced (similar number of instances for each class), with a total of 70 subjects for dysphonia, 82 for laryngitis, 76 for paralysis and 78 healthy subjects. Was applied cross validation [20] on 10 folds to training the models, extracting 10% of the training dataset to be used as validation. A shuffle in the dataset was made before and after each fold is divided into train and test sets. 10% of training set was used to validation set. It was also used early stop condition to stop the training evaluating the validation loss, if the validation loss stops decreasing after five epochs, it stops (this method is called early stopping). The code was developed in Python using the libraries of Keras, Tensorflow, Scikit-learn.

To evaluate the performance it was used Precision, Sensibility, Specificity, F1-Score of the 10 folds Matrix (Global Matrix). Equation 1, 2, 3, 4 shows how to calculate them. Where TP means True Positive, TN means True Negative, FP means False Positive, FN means False Negative and for F1-Score P means Precision and S Sensibility. It was used the classification report function from scikit-learn to obtain the matrix's result. In this function Precision, Sensibility and F1-Score are calculate for each class and it is present the average of those. The average is used as a result. In addition, the Specificity was calculate using Equation 3.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Sensibility = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$F1Score = 2 * \frac{P * S}{P + S} \tag{4}$$

3. Results and discussion

With the implemented models the following 10 folds confusion matrices were obtained, which can be seen in Figure 1, where the left (Figure 1.a) show the matrix of the LSTM model and the right (Figure 1.b) the convolutional 1D. The Table 1 shows the results of the models. The Conv1D model obtained a slight advantage in compare to LSTM with 40% of precision, 41% of sensitivity and 40% of F1-Socore. However, there is still much confusion among pathological classes. The worst case of dysphonia, where the number of instances classified incorrectly is very high. The second worst case is that of Chronic Laryngitis, whose classification also presents many errors, especially in Healthy and Paralysis. The healthy class was the one that had the most correctly classified instances, however, there was a lot of confusion with the laryngitis class, and that really makes sense, because distinguishing this difference even for the ears is difficult.

Table 1. Results for the four classes in LSTM and Conv1D.

Model	Precision %	F1-Score	Sensibility %
LSTM	39	39	40
Conv1D	**40**	**40**	**41**

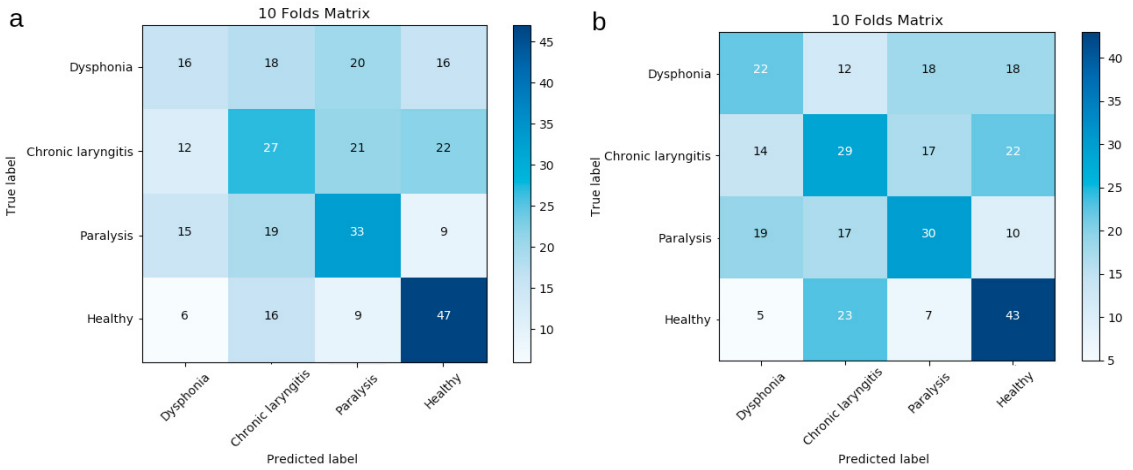

Fig. 1. (a) Model LSTM; (b) Model Conv1D.

Figure 2 shows the performance of the Conv1D model in each fold (10 folds – 10 lines) along the training epochs. Figure 2a shows the training set error and Figure 2.b the validation set error. In this sense, it is possible to visualize that there are cases in which the model presented major and minor validation errors, especially the fold represented by the dark blue line that had the smallest error. The interesting thing is that the errors show a learning potential of the model given the amount of data presented and the errors of training and validation presented. Suggesting the increase of data for better models in the classification of the four classes.

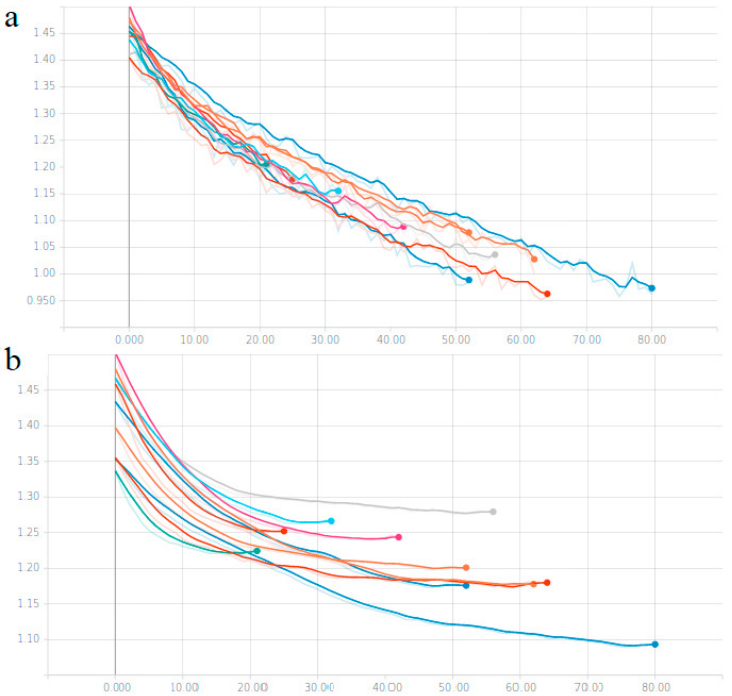

Fig. 2. . (a) Performance of training on each fold; (b) Performance of validation on each fold.

Since the results of the identification of the 4 classes were relatively low power of classification the binary classification was implemented and described below. Three experiments were conducted for the binary classification between each pathology and control (healthy) subjects.

For the binary classification between dysphonia and healthy it is possible to observe in Table 2 that the Conv1D model had a slightly better performance than the LSTM model, obtaining a value of 66% of sensibility, 71% of specificity and 66% in F1-Score, against the 63% sensibility, 69% of specificity and 63% of F1-Score of the second model. The respective class predicted by the networks can be seen in Figure 3. The left (Figure 3.a) representing the LSTM model and the right the Conv1D model (Figure 3.b).

Table 2. Results for Dysphonia x Healthy in LSTM and Conv1D.

Model	Precision %	F1-Score	Sensibility %	Specificity %
LSTM	63	63	63	69
Conv1D	**66**	**66**	**66**	**71**

In these global matrices, the difference between the healthy and dysphonic fit is only two instances, the same for the classes that were classified incorrectly. Thus, despite the results of Table 2, it can be said that both models had close results, both of which could be used for future experiments.

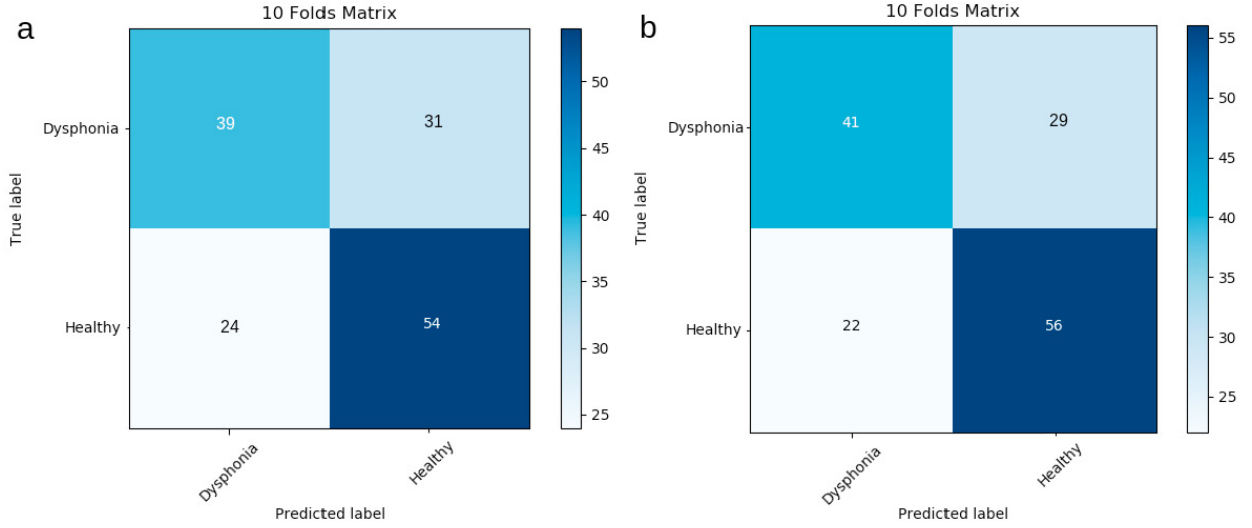

Fig. 3. (a) Model LSTM for Dysphonia x Healthy; (b) Model Conv1D for Dysphonia x Healthy.

The performance for the binary classification between Chronic Laryngitis and healthy is presented in Table 3. The differences between the results were closer, with a slight advantage for the Conv1D model, with values of 67% of sensibility, 67% of specificity and 67% of F1- Score, against 66% of sensibility, 67% of specificity and 66% of F1-Score.

Table 3. Results for Laryngitis x Healthy in LSTM and Conv1D.

Model	Precision %	F1-Score	Sensibility %	Specificity %
LSTM	66	66	66	67
Conv1D	**67**	**67**	**67**	**67**

What made the Conv1D model a little better is that it hit 54 instances of laryngitis against the 52 of the LSTM model. It also rated less erroneous laryngitis data, equivalent to two less-ranked data incorrectly. This difference is show in Figure 4. In this sense, there is no model that has been clearly better.

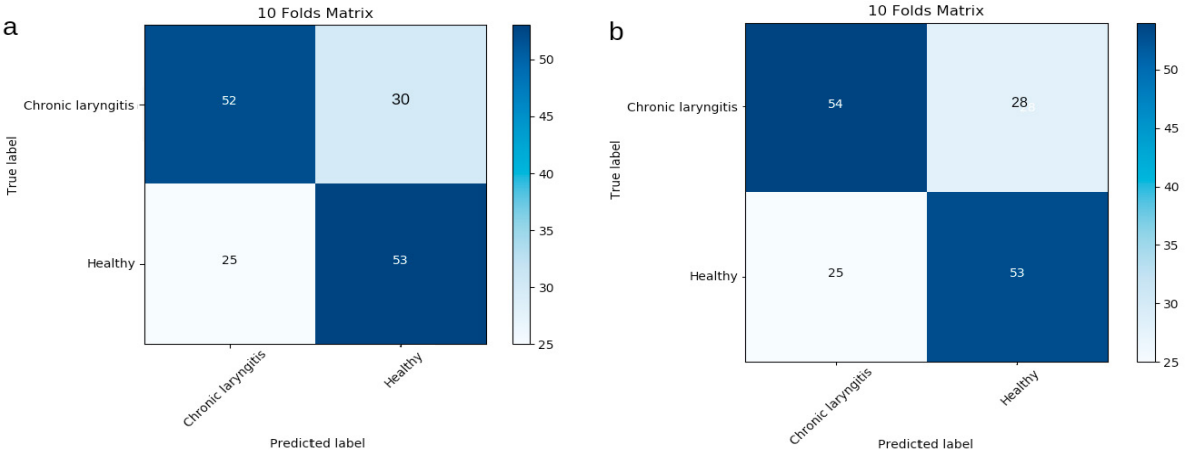

Fig. 4. (a) Model LSTM for Laryngitis x Healthy; (b) Model Conv1D for Laryngitis x Healthy.

The performance analysis for the classifications between paralysis of the vocal cord and healthy subjects, (the dataset balanced with 197 instances) represented by Table 4, it is possible to observe a slight advantage for the LSTM model, obtaining results of 80% for sensibility, specificity and F1-Score, against 78% and 80% of the Conv1D model. This is because the LSTM model hit more data in paralysis, the equivalent of 159 against 150. It also ended up wandering less when the class was paralysis and was classified as healthy. These relationships is shown in Figure 5.

Table 4. Results for Paralysis x Healthy in LSTM and Conv1D.

Model	Precision %	F1-Score	Sensibility %	Specificity %
LSTM	**80**	**80**	**80**	**80**
Conv1D	78	78	78	80

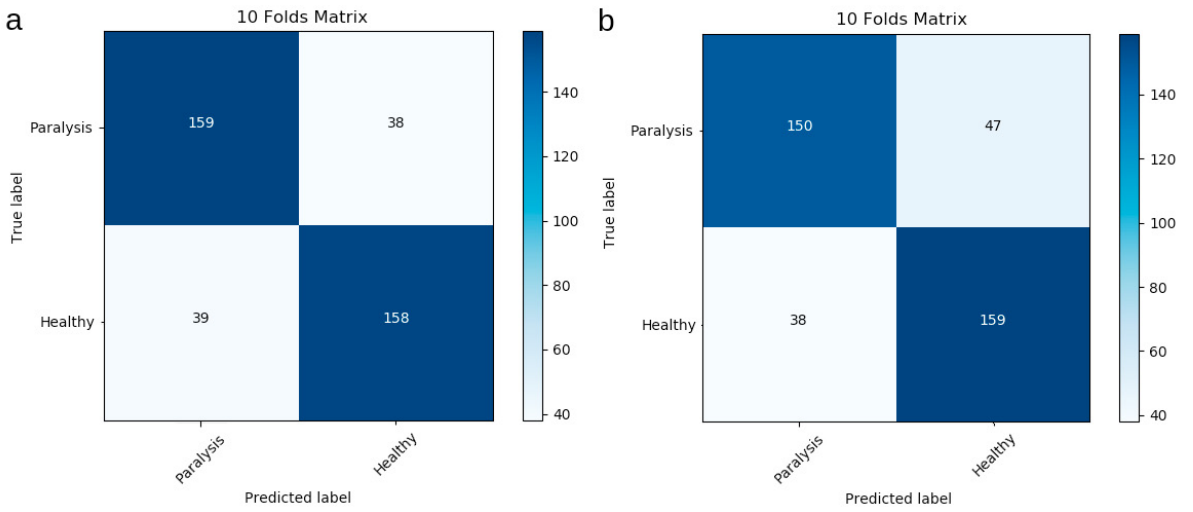

Fig. 5 (a) Model LSTM for Paralysis x Healthy; (b) Model Conv1D for Paralysis x Healthy.

4. Conclusion and Future Work

Therefore, it is noticeable that the use of transfer learning with Google's AudioSet to help in the problem of classification of pathologies is valid as it is possible through the extracted embeddings to carry out such classifications.

However, the classification of speech pathologies through phrases is still a problem that needs to be better studied and addressed by the academic community, mainly due to the amount of cataloged data available.

Concerning the LSTM and Conv1D models presented, it is reasonable to perform such classifications as a binary classification. To classify in multiple classes the models can be taken as promising, however it would be necessary to add more data from each class. It should be considered the hypothesis that using Deep Learning to classify requires a huge amount of data. Another point is that the Conv1D model performed slightly better than the LSTM, but for future analyzes with more data both should be taken into account.

Concerning the binary classification between healthy and pathologic, the vocal cords paralysis, achieve better classification than dysphonia and chronic laryngitis.

The vocal pathologies are harder to classify using continuous speech than using sustained vowels. However, the use of continuous speech may offer applications that are more challenging.

As future work, it would be interesting to create an Open Source software to collect more data for sentences. Preferably with phrases of different languages. In addition, it would also be important to find similar diseases in the German database itself to create sets in the classes used and increase the number of individuals in the analysis.

Acknowledgements

This work is supported by the Fundação para a Ciência e Tecnologia (FCT) under the project number UID/GES/4752/2019.

References

[1] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget and Z. Smekal, (2017) "Voice Pathology Detection Using Deep Learning: a Preliminary Study," 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, 2017, pp. 1-4. doi: 10.1109/IWOBI.2017.7985525. URL: http://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=7985525&isnumber=7985514

[2] M. Alhussein and G. Muhammad, (2018) "Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework," in *IEEE Access*, vol. 6, pp. 41034-41041. doi: 10.1109/ACCESS.2018.2856238
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8411437&isnumber=8274985

[3] Islam, Kazi Aminul & Perez, Daniel & Li, Jiang. (2018). "A Transfer Learning Approach for the 2018 FEMH Voice Data Challenge".

[4] Teixeira Felipe, Fernandes Joana, Guedes Vitor, Junior Arnaldo & Teixeira J. P. (2018). Classification of Control/Pathologic Subjects with Support Vector Machines. Procedia Computer Science. 138. 272-279. 10.1016/j.procs.2018.10.039.

[5] Guedes, V., Junior, A., Teixeira, F., Fernandes, J., & Teixeira, J. P. (2018) "Long Short Term Memory on Chronic Laryngitis Classification", Procedia Computer Science - Elsevier. Volume 138, Pages 250-257.

[6] Teixeira, J. P., Fernandes, P. O. & Alves, N. (2017) "Vocal Acoustic Analysis – Classification of Dysphonic Voices with Artificial Neural Networks", Procedia Computer Science - Elsevier 121, 19–26.

[7] Cordeiro, Hugo & Meneses, Carlos & Fonseca, Jose. (2015). Continuous Speech Classification Systems for Voice Pathologies Identification. IFIP Advances in Information and Communication Technology. 450. 217-224. 10.1007/978-3-319-16766-4_23.

[8] L. Torrey, J. Shavlik (2009). "Transfer learning". IGI Global.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal e M. Ritter, (2017) "Audio set: An ontology and human-labeled dataset for audio events", in Proc. IEEE ICASSP 2017, New Orleans, LA.

[10] S. Hersbey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Saybold, M. Slaney, R. Weiss e K. Wilson, (2017) "Cnn architectures for large-scale audio classification", in International Conference on Acoustics, Speech and Signal Processing (ICASSP). address: https://arxiv.org/abs/1609.09430.

[11] Barry, W.J., Pützer, M.. "Saarbrücken Voice Databse" Institute of Phonetics, Univ. of Saarland. address: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4.

[12] Gröchenig, Karlheinz. (2001). "Foundations of Time--Frequency Analysis". 10.1007/978-1-4612-0003-1.

[13] Jolliffe, I. (2002). "Principal Component Analysis". 2nd edition, Springer, New York.

[14] Gallager, R. (2008). "Principles of Digital Communication". Cambridge Univ. Press, 2008.

[15] Hochreiter, Sepp, and Jürgen Schmidhuber (1997) "Long short-term memory." Neural computation 9.8:1735-1780.

[16] Alex Graves. (2012). "Supervised Sequence Labelling with Recurrent Neural Network". Studies in Computational Intelligence - Springer.

[17] Ioffe, S., & Szegedy, C. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *ICML*.

[18] O'Shea, Keiron & Nash, Ryan. (2015). "An Introduction to Convolutional Neural Networks". ArXiv e-prints.

[19] I. Goodfellow, Y. Bengio e A. Courville. (2016). "Deep learning." MIT Press. Address: http://www.deeplearningbook.org

[20] Refaeilzadeh, P., Tang, L. & Liu, H. (2009). Cross-Validation.. In L. Liu & M. T. Özsu (ed.), Encyclopedia of Database Systems(pp. 532-538) . Springer US . ISBN: 978-0-387-39940-9.