Deep Learning Solution for Pathological Voice Detection using LSTM-based Autoencoder Hybrid with Multi-Task Learning

Dávid Sztahó, Kiss Gábor and Tulics Miklós Gábriel
Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok körútja 2., Budapest, Hungary

Keywords: Deep Neural Networks, Long-Short Term Memory, Autoencoder, Multi-Task Learning, Speech, Bio-signal, LSTM, Voice Pathologies, Parkinson, Depression.

Abstract: In this paper, a deep learning approach is introduced to detect pathological voice disorders from continuous speech. Speech as bio-signal is getting more and more attention as a discriminant for different diseases. To exploit information in speech, a long-short term memory (LSTM) autoencoder hybrid with multi-task learning solution is proposed with spectrogram as input feature. Different speech databases (voice disorders, depression, Parkinson's disease) are applied as evaluation datasets. Applicability of the method is demonstrated by obtaining accuracies 85% for Parkinson's disease, 86% for dysphonia, and 90% for depression on test datasets. The advantage of this method is that it is fully data-driven, in the sense that it does not require special acoustic-phonetic preprocessing separately for the types of disease to be recognized. We believe that the applied method in this article can be used to other diseases as well and can be used for other languages also.

1 INTRODUCTION

Speech as bio-signal getting more and more attention as a discriminant for different diseases. There can be many alterations in speech production due to neurological and/or organic disorders caused by illnesses. In general, any alteration from 'normal' speech might be an indication of pathological speech. Alterations of speech may be caused by various things, for example psychological conditions such as depression. Voice disorders are also main causes of voice alternations. Voice disorder happens once somebody's voice quality, pitch, and loudness are inappropriate for an individual's age, gender, cultural background, or geographic location. The American Speech-Language-Hearing Association divides voice disorders into two groups: organic voice disorders and functional voice disorders. Organic disorders can be structural and neurogenic in nature. Structural disorders involve physical changes in the voice mechanism, such as alterations in vocal fold tissues such as oedema or vocal nodules, polyps, gastroesophageal reflux disease (GERD), cyst and vocal cord paralysis. Neurogenic voice disorders on the other hand are caused by a problem in the nervous system, that include voice problems caused by abnormal control, coordination, or strength of voice box muscles due to an underlying neurological disease such as stroke, Parkinson's disease, multiple sclerosis.

Classifying speech into normal and disordered is more problematic than it first seems. There are a large number of works (Dastjerd et al., 2019; Filiou et al., 2020; Jeancolas et al., 2020; Kiss & Vicsi, 2017a; Klempíř & Krupička, 2018; Low et al., 2020; Tóth et al., 2018; Zhang et al., 2019) subjected to classification of these diseases using various machine learning techniques.

Deep learning (DL) is one of the most frequently used machine learning solutions nowadays. There are many DNN algorithms developed, each is a proper solution to a given data type processing. In this paper, we utilize a long-short term memory (LSTM) autoencoder (AE) hybrid with multi-task learning (MTL) solution to propose a DL structure for detecting multiple diseases, using a voice disorder, a depression and a Parkinson's disease dataset.

An important disadvantage of these classification methods is that they may need complex phonetic preprocessing in order to detect different parts of the speech and, therefore, they may be language dependent. The proposed method in this paper is an

Sztahó, D., Gábor, K. and Gábriel, T.
Deep Learning Solution for Pathological Voice Detection using LSTM-based Autoencoder Hybrid with Multi-Task Learning.
DOI: 10.5220/0010193101350141
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 4: BIOSIGNALS*, pages 135-141
ISBN: 978-989-758-490-9

automatic way without the need of segmenting the speech by an automatic speech recognizer (ASR).

Feature extraction is a critical step of any classification method. DL has the ability to learn to extract the best needed features for the best result. There are end-to-end systems that receive raw sound signals (amplitudes) as input and derive the proper features themselves. For this, huge available data is needed. Generally, dealing with speech and diseases, this is not the case. In our work we use spectrograms as inputs to the DL method and apply autoencoder based feature learning (Yu et al., 2019).

Speech production process is time-varying. The same linguistic content can be said in many durations. Importantly, this variation may not correlate with the given task (disease detection) at all. LSTMs, as a special DL building element, has the property to learn information across time due to its ability to have memory. Therefore, it can learn information that concerns different diseases (Gupta, 2018; Kim et al., 2018; Mallela et al., 2020; Yang et al., 2016; Zhao et al., 2019).

Overfitting is always an important concern in classification trials. There are multiple ways of overcoming this error, mostly by applying proper dataset splitting (train-validation-test). Here, beside the correct dataset splitting, multi-task learning (Ruder, 2004) is also utilized. MTL is not only used as regularization, but also for the parallel classification and autoencoder (feature learning) implementation.

Recent works dealing with deep learning and disease classification include various convolutional network assemblies, recurrent neural networks, LSTMs and even solutions on mobile devices. Since the majority of these works use different datasets for evaluation (even for the same disease), it is hard to compare their results. Most of them report about 90% classification accuracy (Gunduz, 2019; Kaur et al., 2019, 2020; Lam et al., 2019; Mdhaffar et al., 2019; Mohammed et al., 2020; Rejaibi et al., 2019, p.). The proposed AE-LSTM hybrid can be considered as novel architecture among the found studies.

There are several approaches for the binary classification of a healthy subject from voices affected by some disorder. The first question is whether to use sustained vowels or continuous speech. Researchers achieved high accuracies using sustained vowels (Orozco-Arroyave, 2015; Zhang, 2008; Ali, 2017; Teixeira, 2017), however, a significant proportion of researchers use continuous speech in their research pointing out the benefits of using continuous speech over sustained vowels (Vicsi, 2011; Guedes, 2019; Cordeiro, 2015). The research findings are expected to be more applicable to practical work since continuous speech is used in real-world situations. In the work of (Guedes, 2019) the German Saarbrücken Voice Database with the phrase “Guten Morgen, wie geht es Ihnen?” to classify dysphonia and healthy voices. A 66% f1-score was reached in their experiment with Long-Short-Term-Memory and Convolutional Network for classification. In (Tulics, 2019) researchers used acoustic features and phone-level posterior probabilities computed by the DNN soft-max layer of the speech recognition system and used them as an input for an SVM and a Fully-Connected Deep Neural Network. Classification accuracies were ranging from 85% to 88% in their experiments.

The accuracy of distinguishing between depressed and healthy subjects depends largely on the database used, such as the size of the database and the severity of the subjects included in it (Cummins et al., 2015). The accuracy of the classification also depends on the methods used, such as the feature extraction, or the use of gender dependent or independent models (Low et al., 2020). In (Kiss & Vicsi, 2017b) researchers used gender dependent models and used selected acoustic features as an input for an SVM, achieving 86% accuracy with a database which can be considered similar as ours.

The paper is structured as follows: in Section 2, the used speech datasets are described. Section 3 discusses the methods applied. Section 4, 5 and 6 contains the results, discussions and conclusions.

2 DATABASE

The proposed DL structure was tested on three datasets of three disease types. Each dataset contained the recording of reading a short folk tale ‘The North Wind and the Sun’ in Hungarian. For each dataset, healthy speakers are included as a control population with the same sample number as the given disease dataset and age and gender distribution matching the patients’ statistics. Although the classification task described is binary, for the sake of completeness, the severity of the disease is also noted here for each dataset. The audio samples were recorded with external USB sound card and clip-on microphone in PCM format using 44 kHz sampling rate and 16-bit quantization. Informed consents were signed by each patient before recordings.

2.1 Hungarian Parkinson's Speech Dataset (HPSD)

Speech samples were collected from patients diagnosed with Parkinson's disease (PD) by two health institutes in Budapest: Virányos Clinic and Semmelweis University. The severity of PD was labelled according to the Hoehn & Yahr scale (H-Y) (Hoehn & Yahr, 1967). The H-Y scale ranges from 1 to 5, where 1 indicates minimal PD while 5 is the worst PD condition. We did not filter the patients according to the taken medications. All patients were in ON state.

83 speech sample were collected from patients with PD: 43 male speakers (mean H-Y score: 2.74(±1.05); mean age: 64(±9.5)) and 40 female speakers (mean H-Y score: 2.74(±1.10); mean age: 65.4(±9.4)).

2.2 Voice Disorder Speech Dataset (VDSD)

Voice samples from patients were collected during patient consultations in a consulting room at the Department of Head and Neck Surgery of the National Institute of Oncology. The collected speech database contains voices from people suffering from diseases like tumors at various places of the vocal tract, gastroesophageal reflux disease, chronic inflammation of larynx, bulbar paresis, amyotrophic lateral sclerosis, leukoplakia, spasmodic dysphonia, etc. The recorded voice samples in this experiments were classified by a leading phoniatric according to the RBH scale. The RBH scale gives the severity of dysphonia, where R stands for roughness, B for breathiness and H for overall hoarseness. The degree of the category H cannot be less than the highest rate of the other two categories. For example, if B = 3 and R = 2, H is 3, and cannot be 2 or 1. A healthy voice's code is R0B0H0; the maximum H and respectively RBH value is 3, so a voice's code with severe dysphonia is R3B3H3. Here the H score is given.

The database contains a total of 261 recordings from patients with dysphonia: 159 females (mean H score: 1.72(±0.77); mean age: 57.3(±14.8)) and 102 males (mean H score: 2 (±0.83); mean age: 53.7(±15.1)).

2.3 Hungarian Depressed Speech Dataset (HDSD)

Speech samples were collected from patients diagnosed with depression by the Psychiatric and Psychotherapeutic Clinic of Semmelweis University, Budapest. Patients with antipsychotic medication which can affect the acoustical features of speech were left out. The degree of severity of depression was recorded using the Beck Depression Inventory II (BDI) scale (Beck et al., 1996)). The BDI-II scale ranges from 0 to 63, where 0 indicates a healthy state, while 63 is the worst depression condition. The BDI-II scale uses the following rating: 0-13 healthy, 14-19 mild depression, 20-28 moderate depression, 29-63 severe depression.

A total of 107 speech sample were recorded from depressed patients: 64 female subjects (mean BDI score: 28.0(±9.0); mean age: 37.5(±13.5) and 43 males subjects (mean BDI score: 26.1(±8.0); mean age: 40.8(±13.6)).

3 METHODS

3.1 Implemented Deep Learning Architecture and Input

The DL architecture, training and evaluation was implemented in Tensorflow 2.1.0. The implemented DL architecture consists of an LSTM and an autoencoder part. Multi-task learning is applied in order to train the network for the task-specific labels and the autencoder-based feature extraction. Figure 1 shows the structure of the implemented network.

Spectrogram is fed to the network as input. For each audio sample, the spectrogram was extracted by 10 ms timestep and 256 FFT size (resulting in 16 ms window size), commonly used in speech analysis. Because Tensorflow did not manage the varying duration of the samples, this was solved by using a Masking layer at the input. Technically, each spectrogram was padded with 0.0 elements to reach the duration of the longest audio sample. By using the Masking layer, the 0.0 elements were skipped during training and prediction processes.

The DL architecture consists of two parts. An autoencoder part learns feature representation (dimensionality reduction) for the audio sample spectra. This is intended to encode information in the spectra. Part of the bottleneck layer ($fc^{autoencoder}$) has shared neurons that are also trained to the task-specific target labels. This tries to ensure that part of the encoded spectra contains information that is specific to the given disease. This also serves as a regulation technique to avoid overfitting. The idea behind this multi-task learning is that this forces the encoded spectra to contain information partly about the disease characteristics.

The other part of the DL architecture performs disease-specific classification. This part contains two fully connected layers (one is a shared layer in the bottleneck layer) with relu activation functions, dropout layers (dropout parameter set to 0.5) and softmax layer at the end. Before the softmax layer an average pooling layer is needed in order to make one decision for one audio sample (by averaging the outputs of layer fc^t).

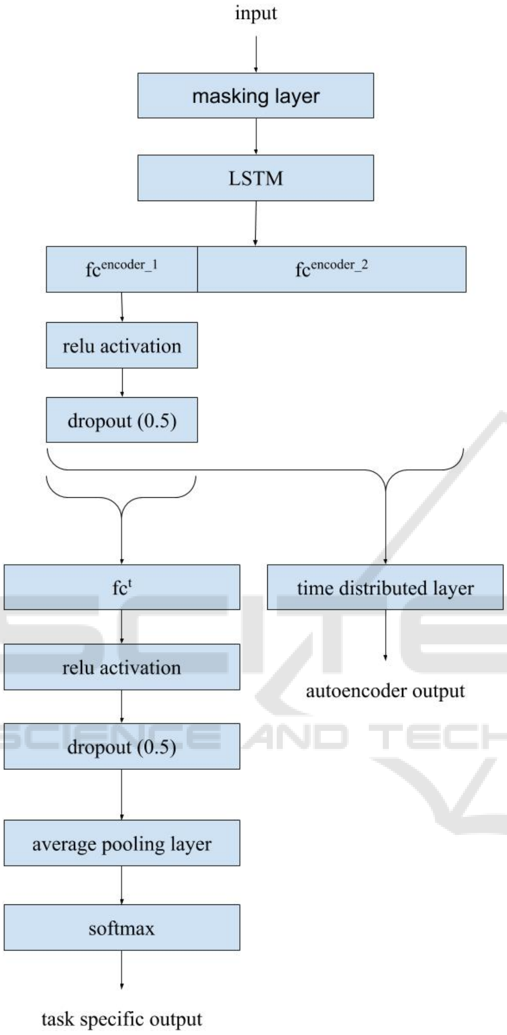

Figure 1: Structure of the implemented network.

Table 1: Number of units in DL layers.

layer name	units
lstm	100
fc^{autoencoder1}	30
fc^{autoencoder2}	30
fc^{t1}	200
time distributed layer	128

A shared LSTM layer is also applied in order to learn time varying information.

Layer sizes are shown in Table 1. The appropriate numbers were selected according to preliminary experiments.

3.2 Dataset Splitting and Network Training

Evaluation was performed by splitting each database into training, validation and test sets by the following method.

Due to the limited number of audio samples available, all samples were used for testing by a 10-fold cross-validation process. 10 test sets were created by 10-fold cross-validation (stratified). In each cross-validation iteration, the remaining 90% of the samples were split into training and validation sets, 70% and 20% respectively, by stratified random sampling. Training was done on the training set and an early stopping was applied on the validation set. Minimum cross-entropy of the disease classification was used as a cost function for early stopping. Maximum 1000 training epochs were done with 50 patience steps for early stopping. During training, 'Adam' optimizer was used.

4 RESULTS

Accuracy, sensitivity and specificity was used as evaluation metrics. Tables 2, 3 and 4 show the results in each set during tests (training, validation and testing). The confusion matrices for the test sets are also shown in Table 5. The values in the cells are the number of samples (subjects).

The results show that the applied DL performs well on all three datasets. The lowest accuracy on the test sets is 0.86, which means that 86% of the samples are correctly classified into healthy or disease categories. The highest score is 0.90 in the case of depression.

The DL method doesn't seem to overfit. Balanced results metrics are achieved in the training, validation and test steps.

Also, sensitivity and specificity scores are well balanced in the case of VDSD and HDSD datasets. Parkinson samples are a little bit unbalanced according to these metrics. Higher specificity is reached, which means that healthy samples are classified more accurate than the Parkinson samples.

Table 2: Results on the HPDS samples.

	accuracy	sensitivity	specificity
training	0.86	0.77	0.95
validation	0.88	0.80	0.96
test	0.85	0.75	0.95

Table 3: Results on the VDSD samples.

	accuracy	sensitivity	specificity
training	0.90	0.89	0.91
validation	0.87	0.86	0.89
test	0.86	0.85	0.87

Table 4: Results on the HDSD samples.

	accuracy	sensitivity	specificity
training	0.93	0.92	0.93
validation	0.90	0.90	0.89
test	0.90	0.91	0.89

Table 5: Confusion matrices obtained on the test datasets.

PDSD		
	predicted positive	predicted negative
true positive	62	21
true negative	4	79
PDSD		
	predicted positive	predicted negative
true positive	226	35
true negative	28	164
HDSD		
	predicted positive	predicted negative
true positive	95	12
true negative	9	93

5 DISCUSSION

The achieved results on the applied three datasets show that the proposed DL method is able to extract information from the samples in order to make distinction between negative (healthy) and positive cases. These actual accuracy, sensitivity and specificity scores are, naturally, dependent on the datasets. However, they can be considered large enough, that some statements could be concluded based on them. With the extension of the recordings and by applying more datasets in different languages may also prove the proposed DL method to be more robustly usable.

Based on the results, the highest evaluation scores were achieved on the depression database test set (and also on training and validation sets). Based on personal experience of clinical experts, intonation is also highly affected by depression. By applying an LSTM layer, this intonational disorder (which can be captured through temporal analysis) can be more accurately modelled.

The suggested method not only captured intonational features of speech, but voice characteristics related to dysphonia, such as hoarseness and breathiness.

In case of Parkinson's disease lower sensitivity scores are achieved than specificity scores. Although a high sensitivity-specificity balance is more desirable, the present case doesn't mean that the method can't be used as pre-screening. Less positive samples will be detected, but the overall accuracy can be considered sufficient for the task. In fact, every method is usable with over random performance.

The results achieved are comparable to the previous results in the field. Since actual results are dataset dependent, direct comparison of accuracy and other metrics is problematic. (Kiss & Vicsi, 2017b) reported 86% accuracy for a former version of the depression dataset. Here, we achieved 90%. In case of PD, (Sztahó et al., 2019) reported around 88% accuracy using cross-validation setup, without separate independent test set. The 85% achieved here is comparable, especially if we add that here we applied an appropriate test set. In the case of VDSD dataset, a previous result is reported in (Tulics, 2019). In that, a 95% accuracy was described with possible overfitting effect, and between 85-88% without overfitting. Here, we achieved 86% without probable overfitting. All these researches used segmentation information to obtain the highest performance. In our case here, this computationally intensive step is not needed.

Among many other usages, actual practical applicability can be pre-screening in general practitioner offices or home-care environments. A cheap, easy to use devices (software) can be implemented to detect various diseases that affect speech.

6 CONCLUSION

The goal of this work is to introduce a novel method for pathological speech recognition using continuous speech without the need of a voicing detection or speech segmentation application. The proposed DL architecture consists of two parts: an autoencoder part learns feature representation and a disease-specific classification.

We demonstrated the applicability of the method by classifying three different diseases: Parkinson's disease, dysphonia related voice disorders and depression. The method achieved 0.85 for Parkinson's disease, 0.86 for dysphonia, and 0.90 for depression on the test datasets. These classification accuracies correspond to the classification accuracies mentioned in the literature. The advantage of this method is that it is fully data-driven, in the sense that it does not require special acoustic-phonetic preprocessing separately for the types of disease to be recognized. The speech recordings can be directly given to the deep neural network (using spectrographic extraction only).

We believe that the applied method in this article can be used to other diseases as well and can be used for other languages also.

ACKNOWLEDGEMENTS

Project no. K128568 has been implemented with the support provided from 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

Ali, Z., Talha, M., & Alsulaiman, M., 2017. A practical approach: Design and implementation of a healthcare software for screening of dysphonic patients. *IEEE Access*, 5, 5844-5857.

Beck, A. T., Steer, R. A., Ball, R. & Ranieri, W. F., 1996. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment* 67, 588–597.

Cordeiro, H., Meneses, C., & Fonseca, J., 2015. Continuous speech classification systems for voice pathologies identification. In *Doctoral Conference on Computing, Electrical and Industrial Systems*, pp. 217-224.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F., 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71, 10-49.

Dastjerd, N. K., Sert, O. C., Ozyer, T., & Alhajj, R., 2019. Fuzzy Classification Methods Based Diagnosis of Parkinson's disease from Speech Test Cases. *Current Aging Science*, 12(2), 100–120.

Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., & Brambati, S. M., 2020. Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: A scoping review. *Aphasiology*, 34(6), 723–755.

Guedes, V., Teixeira, F., Oliveira, A., Fernandes, J., Silva, L., Junior, A., & Teixeira, J. P., 2019. Transfer Learning with AudioSet to Voice Pathologies Identification in Continuous Speech. *Procedia Computer Science*, 164, 662-669.

Gunduz, H., 2019. Deep learning-based Parkinson's disease classification using vocal feature sets. *IEEE Access*, 7, 115540–115551.

Gupta, V., 2018. Voice disorder detection using long short term memory (lstm) model. ArXiv Preprint ArXiv:1812.01779.

Hoehn, M. & Yahr, M. D., 1967. Parkinsonism onset, progression, and mortality. *Neurology* 17, pp. 427–427

Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B.-E., Corvol, J.-C., Vidailhet, M., Lehéricy, S., & Benali, H., 2020. X-vectors: New Quantitative Biomarkers for Early Parkinson's Disease Detection from Speech. ArXiv:2007.03599 [Cs, Eess, q-Bio]. http://arxiv.org/abs/2007.03599

Kaur, S., Aggarwal, H., & Rani, R., 2019. Diagnosis of Parkinson's Disease Using Principle Component Analysis and Deep Learning. *Journal of Medical Imaging and Health Informatics*, 9(3), 602–609.

Kaur, S., Aggarwal, H., & Rani, R., 2020. Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. *Machine Vision and Applications*, 31(5), 32.

Kim, M. J., Cao, B., An, K., & Wang, J., 2018. Dysarthric Speech Recognition Using Convolutional LSTM Neural Network. *INTERSPEECH*, 2948–2952.

Kiss, G., & Vicsi, K., 2017a. Comparison of read and spontaneous speech in case of automatic detection of depression. 2017 *8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 213–218.

Kiss, G., & Vicsi, K., 2017b. Mono-and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology*, 20(4), 919-935.

Klempíř, O., & Krupička, R., 2018. Machine learning using speech utterances for Parkinson disease detection. Lékař a Technika - *Clinician and Technology*, 48(2), 66–71.

Lam, G., Dongyan, H., & Lin, W., 2019. Context-aware deep learning for multi-modal depression detection. ICASSP 2019-2019 *IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 3946–3950.

Low, D. M., Bentley, K. H., & Ghosh, S. S., 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96–116.

Mallela, J., Illa, A., Suhas, B. N., Udupa, S., Belur, Y., Atchayaram, N., Yadav, R., Reddy, P., Gope, D., & Ghosh, P. K., 2020. Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and Healthy Controls with CNN-LSTM using transfer learning. ICASSP 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6784–6788.

Mdhaffar, A., Cherif, F., Kessentini, Y., Maalej, M., Thabet, J. B., Maalej, M., Jmaiel, M., & Freisleben, B., 2019. DL4DED: Deep Learning for Depressive Episode Detection on Mobile Devices. *International Conference on Smart Homes and Health Telematics*, 109–121.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., & AL-Dhief, F. T., 2020. Voice Pathology Detection and Classification Using Convolutional Neural Network Model. *Applied Sciences*, 10(11), 3723.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., & Nöth, E., 2015. Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A., 2019. MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech. ArXiv Preprint ArXiv:1909.07208.

Ruder, S., 2017. An overview of multi-task learning in deep neural networks. ArXiv Preprint ArXiv:1706.05098.

Sztahó, D., Valálik, I., & Vicsi, K., 2019. Parkinson's Disease Severity Estimation on Hungarian Speech Using Various Speech Tasks. In *International Conference on Speech Technology and Human-Computer Dialogue* (SpeD). pp. 1-6.

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z. & Kálmán, J., 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2), 130-138.

Tulics, M. G., Szaszák, G., Mészáros, K., & Vicsi, K., 2019. Artificial Neural Network and SVM based Voice Disorder Classification. In *10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. pp. 307-312.

Vicsi, K., Imre, V., & Mészáros, K., 2011. Voice disorder detection on the basis of continuous speech. In *5th European Conference of the International Federation for Medical and Biological Engineering*, pp. 86-89.

Yang, T.-H., Wu, C.-H., Huang, K.-Y., & Su, M.-H., 2016. Detection of mood disorder using speech emotion profiles and LSTM. *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5.

Yu, J., Zheng, X., & Wang, S., 2019. A deep autoencoder feature learning method for process pattern recognition. *Journal of Process Control*, 79, pp. 1–15.

Zhang, H., Song, C., Wang, A., Xu, C., Li, D., & Xu, W., 2019. PDVocal: Towards Privacy-preserving Parkinson's Disease Detection using Non-speech Body Sounds. *The 25th Annual International Conference on Mobile Computing and Networking*, pp. 1–16.

Zhang, Y., & Jiang, J. J., 2008. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*, 22(1), pp. 1-9.

Zhao, J., Mao, X., & Chen, L., 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, pp. 312–323.