DOMAIN ADVERSARIAL CONVOLUTIONAL NEURAL NETWORK FOR PARKINSON'S DISEASE DETECTION FROM SPEECH

E. J. Ibarra-Sulbaran[1], J. D. Arias-Londoño [2], M. Zañartu[1], J. I. Godino-Llorente[2]

[1] Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile.
[2] Bioengineering and Optoelectronics lab (ByO), Universidad Politécnica de Madrid, Madrid, Spain
emiro.ibarra@sansano.usm.cl; julian.arias@upm.es; matias.zanartu@usm.cl; ignacio.godino@upm.es

Abstract: **Deep learning has gained popularity in detecting Parkinson's disease (PD) from speech due to its ability to automatically extract meaningful representations from raw data. The most popular approaches are based on Convolutional Neural Network (CNN) models fed with spectrograms. However, the use of these algorithms is constrained due to the cross-dataset accuracy obtained during the validation process. Thus, in this work, we focus on studying the cross-domain effect -specifically due to different databases- for the screening of PD using a CNN-based model and two different speech corpora. To address the cross-domain challenge, we propose the use of domain adversarial (DA) training as a method to obtain discriminant and domain-invariant models. The visualization of the feature space distribution extracted by this model, using t-distributed stochastic neighbor embeddings, along with its divergence and variance by class, indicates a significant improvement in domain adaptation. These initial results provide valuable insights for further model refinement and constitute a proof of concept that domain adversarial methods offer a feasible option for creating a more generalizable speech-based PD detection model.**

Keywords: **Convolutional Neural Networks, Deep learning, Domain Adversarial, Parkinson's Disease.**

I. INTRODUCTION

Several studies have explored end-to-end deep learning techniques for screening PD directly from raw speech and time–frequency spectrograms. These techniques include CNNs [1-4], recurrent neural networks (RNNs) [5], long short-term memory (LSTM) models [6], and others [1]. Among them, CNNs have emerged as the most popular technique.

Most of the reported end-to-end deep learning methods have demonstrated high discriminative capacity in distinguishing between healthy controls (HC) and PD compared to traditional machine learning approaches. However, it is worth noting that the training and validation processes of these algorithms have been developed using a single domain, meaning a single corpus with participants sharing similar demographics, dialectal and recording conditions.

In this line, [1, 2, 4] reveal the limitations of these models for the screening of PD when they are applied to a new dataset, resulting in a drop of accuracy of 20 absolute points. This fact highlights a significant limitation of current methods, demonstrating a noticeable degradation in their discriminative capabilities across domains. Additionally, the model relies on shortcut learning when possible, meaning it learns characteristics that differentiate between the groups but do not generalize well with respect to the underlying pathology.

In this context, we propose adding a domain adaptation step into the representation learning process, which would help to reduce the existing gap between different corpora. The goal is to ensure that the automatic screening of PD is based on features that are both discriminative and invariant to dataset changes.

In the deep learning literature, we came across the domain adversarial training method proposed in [7]. This method suggests an adversarial framework to learn domain-invariant representations. Recently, [8] proposed a speech PD classification using adversarial training to obtain speaker identity-invariant representations within a single corpus. However, they do not consider the effect of multi-dataset scenarios.

The contributions of this work are twofold: First, to analyze the robustness of an end-to-end deep learning method for PD diagnosis with respect to the shift between domains (different speech databases). Second, to study the capacity of domain adversarial training in providing more generic and reliable models for the automatic screening of PD from the speech, addressing undesired speech recording variability.

II. MATERIALS AND METHODS

In this preliminary study, we establish a baseline model to detect PD from Mel-spectrograms by combining a CNN and a multi-layer perceptron (MLP) network, similar to those evaluated in [2-3]. We use two speech corpora to train and test the baseline model, first in a cross-domain test and then by mixing both datasets. Subsequently, the baseline model is adapted using a domain adversarial approach.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudia Manfredi (edited by), *Models and Analysis of Vocal Emissions for Biomedical Applications. 13th International Workshop, September, 12-13, 2023*,
© 2023 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0146-9, DOI 10.36253/979-12-215-0146-9

A. Speech Corpora

The datasets used in this work were previously reported as Gita [9] and Neurovoz [10].

The Gita dataset was recorded by Clínica Noel in Medellín, Colombia. This dataset includes, among other data, diadochokinetic (DDK) tasks (i.e., repetitions of the syllable sequence /pa-ta-ka/) from 100 Colombian Spanish native speakers, with 50 HC and 50 PD patients.

The Neurovoz dataset was collected by Universidad Politécnica de Madrid in collaboration with Gregorio Marañón Hospital in Madrid, Spain. This dataset includes, among other material, DDK sequences from 86 adult speakers whose mother tongue is Castilian Spanish (44 HC and 42 PD).

Recordings for both corpora were obtained under controlled ambient conditions using a sampling rate of 44.1 kHz and 16 bits of quantization. Both datasets were recorded in compliance with the Helsinki Declaration and approved by their respective Ethics Committee.

B. Method

The DDK speech recordings were first normalized using the maximum absolute value of amplitude. They were then segmented into 400 ms intervals overlapped 50%. Each segment was transformed into a time-frequency representation using Mel-scale spectrograms with a window size of 15 ms, a hop length of 10 ms, and 65 Mel bands. This pre-processing resulted in Mel-spectrograms of 65x41 points, which were individually normalized following a Z-score.

The baseline model consists of two modules, which we have named the feature generation network and the PD prediction network. The feature generation network receives Mel-spectrograms as input. This first module is composed of a two-dimensional convolutional layer, where each convolutional layer is followed by a batch normalization, a ReLU activation function, max-pooling (filter size: 3×3), and a dropout layer. Subsequently, the dynamic features obtained from the feature generation network are flattened to connect with the PD predictor network. This second module consists of two fully connected layers with a dropout layer in between to regularize the weights. ReLU activation is used in the first hidden layers, and a SoftMax activation function is used for classification.

For domain adversarial training, the baseline model is adapted following the Domain-Adversarial Neural Network proposed in [7]. This is accomplished by attaching a domain predictor network to the feature extractor network via a Gradient Reversal Layer (GRL). This new module contains the same architecture as the PD prediction network. The only non-standard component of the domain adversarial architecture is the GRL, which leaves the input unchanged during forward propagation and reverses the gradient by multiplying it by a negative scalar during backpropagation [7]. The gradient reversal ensures that the feature distributions over the two domains are as indistinguishable as possible for the domain classifier, providing domain-invariant features.

Regarding training and evaluation, a stratified speaker-independent 10-fold cross-validation was used, ensuring no overlap of speakers across different folds. The hyperparameters of the baseline model were tuned with the 10-fold set of mixed data (Gita and Neurovoz) using Talos [11]. The hyperparameter search space is summarized in Table 1. The model with the best performance on the validation set for the 10 folds was selected for all experiments, including domain adversarial training (DA), where the domain predictor network parameters were set to the same values as the PD prediction network parameters.

The models were trained using the Stochastic Gradient Descent (SGD) algorithm with cross-entropy as the loss function. A learning rate schedule was used, initialized at 0.1. The PyTorch implementation of our approach is available online[1].

Table 1. *Hyperparameters search space for the baseline model*

Hyperparameter	values
Training Batch size	16, 32, 64
Kernel size of conv. layer I	4, 6, 8
Kernel size of conv. layer II	5, 7, 9
Dropout rate	0.2, 0.5
Depth of convolutional layers	32, 64, 128
Units of each fully connected layer	16, 32, 64

III. RESULTS

For mixed data training, the features extracted from the last layer of the PD prediction network for each model were labelled by class (PD and HC) and domain (Gita and Neurovoz). These features were visualized in a two-dimensional map using t-distributed stochastic neighbor embeddings (t-SNE) to study the domain adaptation effect of the baseline model in comparison to the domain adversarial network. A divergence measure was used to quantify the differences in the distribution of domain-labelled features for each class. This measure is computed using the Kullback-Leibler algorithm proposed in [12]. Additionally, the trace of the covariance matrix of the features for each class is used to quantify its variability.

A. Cross-Domain Results

Table 2 shows the validation results obtained for the

[1] https://github.com/Emiroji/Domain_Adversarial_CNN_Speech_Parkinson_Clasification

baseline model trained using individual datasets. The accuracy, sensitivity, specificity, and area under the ROC curve obtained for the validation sets for both Gita and Neurovoz are consistent with those reported in previous work [1-3]. We emphasize the accuracy difference between the validation and cross-domain test, which is over 30 and 20 absolute points for Gita and Neurovoz respectively. This drop in accuracy is aligned with what has been reported in the literature [1, 2, 4], confirming the mentioned limitation of end-to-end approaches trained with a limited dataset.

Table 2. *Classification with the baseline model for each corpus. Acc: accuracy. Sens: Sensitivity. Spec: Specificity. AUC: Area under the ROC curve. Values represent the mean of 10-folds ± standard deviation.*

	Gita	**Neurovoz**
Acc. (%)	80.8 ± 13.4	80.1 ± 16.7
Sens. (%)	81.8	80.5
Spec. (%)	80.0	81.00
AUC	0.9	0.9
Cross-Domain Acc. (%)	**47.7 ± 3.5**	**56. 3 ± 2.6**

B. Domain Adversarial results.

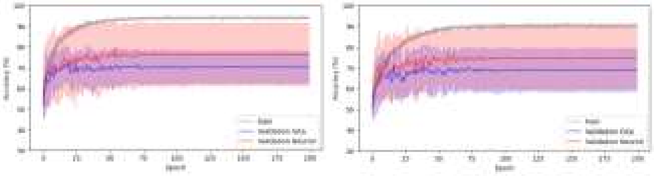
Table 3 contrasts the results obtained by the baseline model and the domain adversarial network, both trained using the mixed speech corpora. The mean validation metrics decrease in comparison with experiment one (where only one dataset is used in the training process), especially for the baseline model. For example, the accuracy in the baseline model dropped by almost 8 and 4 absolute points for Gita and Neurovoz, respectively, whereas for the DA model, it was less than 5 absolute points for both validation sets, with a standard deviation slightly lower for the adversarial scheme.

It is important to highlight that the validation metrics in Table 3 are computed based on estimations by subject. Therefore, the training and validation accuracies in terms of samples with respect to the epochs are shown in Figure 1. From these learning curves, we observe that the models reach stability before 100 epochs. As expected, the small size of the training dataset leads to overfitting of these deep learning models. These curves are consistent with those shown in [1].

The most relevant result obtained in this work is shown in Figure 2. The t-SNE representations show the features extracted by the baseline model and domain adversarial models for the fold with the best validation accuracy (the t-SNE representations for the remaining folds are available in the online GitHub repository[1]). The features extracted by the baseline model in the training set (Figure 2.a) report more than two classification clusters. In contrast, the domain adversarial model shows that the features of the PD cluster for both Gita and Neurovoz share the same space, as well as for HC (see Figure 2.b). A similar trend is observed in the validation set (Figure 2.c and 2.d). However, as expected, this behavior is affected by the model's classification performance, being more evident in those folds where the model shows high accuracy.

Table 3. *Classification with the baseline and DA models for mixed speech corpora. Acc: Accuracy. Sens: Sensitivity. Spec: Specificity. AUC: Area under the ROC curve. Values represent the mean of 10-folds ± standard deviation.*

	Baseline model		**DA Model**	
	Gita	**Neurovoz**	**Gita**	**Neurovoz**
Acc.	71.9±8.8	76.7±21.6	76.1±11.8	80.6±19.6
Sens.	70.5	80.0	76.5	80.5
Spec.	74.0	73.5	76.0	81.0
AUC	0.9±0.1	0.9±0.2	0.8±0.2	0.9±0.2



(a) (b)

Fig. 1. *Accuracy learning Curves during the k-fold cross-validation: (a) Baseline Model; (b) DA model. The solid line represents the mean values and the shaded regions standard deviation.*

On the other hand, Table 4 shows that both the divergence and the trace of the covariance matrix between domains for each class are higher for the baseline model in contrast to its DA version. The high divergence shown in both HC and PD classes for the baseline model indicates that it presents a higher dissimilarity between the feature distributions extracted for Gita and Neurovoz. On the other hand, the results of the trace of the covariance matrix show that the baseline model exhibits a higher spread for each class.

IV. Discussion and Conclusions

In this study, domain adaptation of an end-to-end CNN-based model for automatically PD detection using a DDK was analyzed. Although most recent work continues to compare models based solely on their accuracy in a single database, this work provides new evidence that these approaches require domain adaptation strategies to be more generalizable

The first experiment showed that the CNN-based model learns characteristics of PD speech during internal validation. However, when tested on unseen datasets, the model failed to identify PD with sufficient accuracy. Subsequently, when both datasets were mixed during training, the features learned by the baseline model for a specific class presented a different distribution. This behavior indicates that the model is extracting additional information relative to domain variability instead of solely obtaining PD discriminative features.

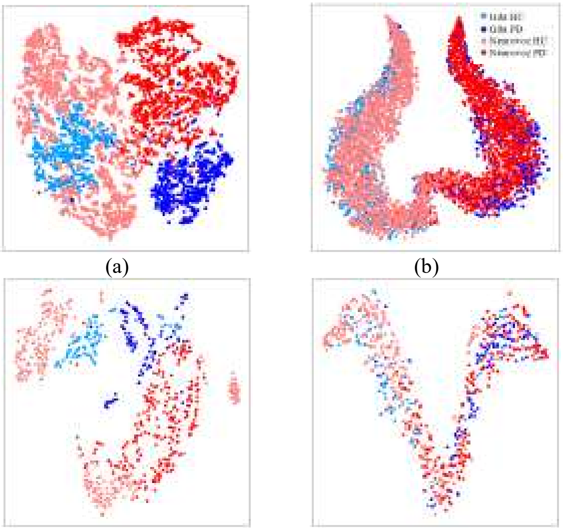

(a) (b)
(c) (d)

Fig. 2. *T-SNE of the extracted features for the training set: (a) baseline model; (b) DA model. And for the validation set: (c) baseline model; (d) DA model.*

Table 4. *Divergence and trace of the covariance matrix between domain distributions for each class. Values represent the mean of 10-folds ± standard deviation.*

	Divergence		**Variance**	
	HC	**PD**	**HC**	**PD**
Baseline	48.8±13.5	50.6±9,9	33.2±15.0	17.1±4.5
DA	15.5±9.3	13.5±5.5	11.3±2.2	9.8±2.9

In contrast, domain adversarial training ensures that the model learns invariant domain features. This is evidenced by the t-SNE visualizations and by the divergence and variance metrics. These preliminary results suggest that domain adversarial training improves the generalization abilities of the network. Nevertheless, more experiments, including new speech corpora, different phonation tasks, and new architectural models, are needed.

VI. Acknowledgements

E.J.I. thanks the support of Agencia Nacional de Investigación y Desarrollo (ANID), "Beca Doctorado Nacional 21190074". This work was also supported by the Ministry of Economy & Competitiveness, Spain, under grants DPI2017-83405-R1 and PID2021-128469OB-I00.

REFERENCES

[1] C. Quan, K. Ren, Z. Luo, Z. Chen, Y. Ling, "End-to-end deep learning approach for Parkinson's disease detection from speech signals," Biocybern Biomed Eng, vol. 42, no. 2, pp. 556–574, 2022.

[2] J. Vásquez-Correa, J. R. Orozco-Arroyave, E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in Proc. Interspeech 2017, pp. 314–318.

[3] J. C. Vásquez-Correa, C. D. Rios-Urrego, T. Arias-Vergara, M. Schuster, J. Rusz, E. Nöth, J. R. Orozco-Arroyave, "Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages," Pattern Recognit. Lett., vol. 150, pp. 272–279, 2021.

[4] Hireš, M., Drotár, P., Pah, NN., Ngo, Q., Kumar, D., "Strengths and Limitations of Computerized PD Diagnosis from Voice". Available at SSRN: https://ssrn.com/abstract=4327662, 2023.

[5] T. Fujita, Z. Luo, C. Quan, K. Mori, S. Cao, "Performance evaluation of RNN with hyperbolic secant in gate structure through application of Parkinson's disease detection," Appl. Sci., vol. 11, no. 10, 2021.

[6] B. E. Mehmet, I. Esme, I. Ibrahim, "Parkinson's detection based on combined CNN and LSTM using enhanced speech signals with variational mode decomposition," Biomed Signal Process Control, vol. 70, p. 103006, 2021.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach Learn Res., vol. 17, no. 59, pp. 1–35, 2016.

[8] P. Janbakhshi, I. Kodrasi, "Supervised speech representation learning for Parkinson's disease classification," Proc. ITG Conf. on Speech Communication, July 2021.

[9] J. Orozco, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, E. Nöth, "New spanish speech corpus database for the analysis of people suffering from Parkinson`s disease," Proc. 9th Lang. Resources and Evaluation Conf. (LREC), 2014, pp. 342–347.

[10] L. Moro-Velazquez, J. Gomez-Garcia, J. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, N. Dehak, "A forced gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing," Biomed Signal Process Control, vol. 48, pp. 205–220, 2019.

[11] Autonomio Talos [Computer software]. (2020). Retrieved from http://github.com/autonomio/talos. Accessed on: 26 July. 2023.

[12] F. Perez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," 2008 IEEE Int. Symposium on Information Theory, Toronto, ON, Canada, 2008, pp. 1666-1670.