DOI: 10.17469/O2111AISV000007

CLAUDIO FERRANTE, VINCENZO SCOTTI

Cross-Lingual Transferability of Voice Analysis Models: a Parkinson's Disease Case Study

Traditionally, speech analysis has always relied on a set of very informative features like (Mel) spectrogram, Mel Frequency Cepstral Coefficients (MFCC), pitch or intensity to build speech powered applications. Recently, deep learning-based models for the extraction of acoustic features have allowed significantly improving the state of the art in many speech-related applications. With this work, we focus the analysis on the cross-lingual transferability of speech analysis features. The idea is to understand whether and how well a classification model trained on speech features in a source language works on an unseen target language. We evaluate these properties analysing models for Parkinson's disease detection from speech, adapting the models from English to Telugu. Results show that multi-lingual pre-trained deep learning-based features do not require explicit adaptation and work well out-of-the-box. Differently, models not adapting out-of-the-box respond well even to unsupervised adaptation on a small data set.

Keywords: Natural Language Processing, Deep Learning, Speech Analysis, Parkinson's Disease, Domain Adaptation, Multilingual

1. *Introduction*

Deep learning has become pervasive in Natural Language Processing (NLP), enabling complex language-based applications. Deep neural network models for language and speech analysis are now capable of complex processing like dialogue generation (Gao, Galley, and Li, 2019) and expressive speech synthesis (Tan, Qin, Soong and Liu, 2021). Now the focus is shifting towards multi-task and multi-language support, enabled by data availability and computing power (Raffel. Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li and Liu, 2020; Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever and Amodei, 2020).

These deep learning-based models are powerful end-to-end tools to solve language-related tasks and can also be used for feature extraction. The hidden transformations learnt by these deep learning models can be reused on other tasks involving input data from a similar domain through the transfer learning process. This process consists in reusing the features extracted by the hidden transformations of a pre-trained deep neural network as input to a new classification neural network (Yosinski, Clune, Bengio and Lipson, 2014).

In this paper, we focus on analysing speech features and their cross-lingual transferability. The objective is to evaluate how well a model trained in one language (hereafter referred to as source language), starting from deep learning-based features, can be reused in another language (hereafter referred to as target language) and whether it is necessary to adopt explicit measures to ensure the adaptation goes fine. The objective is to analyse the generality of these features. We selected Parkinson's disease recognition from speech as the target task, and we used English and Telugu, respectively, as source and target languages.

We divide the rest of this paper into the following sections. In § 2 we present the research work connected to the one we are proposing. In § 3 we introduce the neural network-based classification pipeline we adopted. In § 4 we present the available data, as well as the pre-processing and preparation steps. In § 5 and § 6 we describe the experiments we conducted and the results we obtained. Finally, in § 7 we summarise our work and present hints of possible future works.

2. *Related works*

In the following, we introduce the deep neural network models for acoustic features extraction, the existing machine learning-based solutions for Parkinson's disease detection and the concepts behind domain adaptation.

2.1 Neural networks for speech analysis

Traditionally, speech-based classification and recognition tasks were solved using handcrafted prosodic and acoustic features like (Mel-)spectrogram, Mel Frequency Cepstral Coefficients (MFCC), energy, or pitch (Giannakopoulos 2015; Chernykh, Sterling and Prihodko, 2017). Nowadays, deep learning-based features are slightly substituting these handcrafted ones, thanks to the improved performances they are enabling. Usually, learning good deep learning-based features requires large data sets, so this approach is not directly applicable to any task. However, once trained, the feature extraction modules can be reused effectively also on problems with limited data availability. There exist different pre-trained audio and speech analysis models that can be transferred on other tasks.

SoundNet (Aytar Vondrick and Torralba, 2016) is a 1D convolutional neural network for raw audio analysis. The original neural network was trained to mimic the output of video analysis models for object and scene recognition starting from the audio track. The reference vision model would analyse the images from a video to provide the pseudo-labels for classification, and SoundNet would process the audio learning to predict the pseudo-labels from it. The hidden features have been adapted successfully to other classification tasks, like highlight detection in sports events (Merler, Joshi, Mac, Nguyen, Hammer, Kent, Xiong, Do, Smith, and Schmidt Feris, 2018).

VGGish (Hershey, Chaudhuri, Ellis, Gemmeke, Jansen, Moore, Plakal, Platt, Saurous, Seybold, Slaney, Weiss, and Wilson, 2017) is a 2D convolutional neural

network obtained by refining the VGG neural network for image classification. Differently from SoundNet, this model takes as input the audio spectrograms, interpreting them as greyscale images, instead of raw audio waveforms. The model, in this case, was pre-trained on a large audio classification task. As for SoundNet, the hidden features of VGGish have been adapted to other classification tasks like emotion recognition from speech (Scotti, Galati, Sbattella, and Tedesco, 2020).

Wav2Vec and Wav2Vec 2.0 (Schneider, Baevski, Collobert, and Auli, 2019; Baevski, Zhou, Mohamed, and Auli, 2020) differ from the previous two models in training approach: they adopt self-supervised generative training rather than supervised training. This training approach does not require any labelling, allowing to scale the data set size significantly compared to the other approaches. The models are trained to reconstruct the input speech signal from a compressed representation. The second version adopts a state-of-the-art transformer architecture, and both models were developed initially to be refined for Automatic Speech Recognition (ASR).

2.2 Machine learning for Parkinson's disease detection

The approach to Parkison's disease recognition from speech has evolved following the advances of NLP. Initial recognition pipelines used handcrafted features as input and different machine leaning-based algorithms to solve the task, while currently, the approach is to transfer deep learning-based features.

Initial solutions for Parkison's disease detection using speech analysis and machine explored a wide variety of feature combinations and different classification algorithms (Williamson, Quatieri, Helfer, Perricone, Ghosh, Ciccarelli, and Mehta, 2015; Dimauro, Caivano, Bevilacqua, Girardi, and Napoletano, 2016; Dimauro, Di Nicola, Bevilacqua, Caivano, and Girardi 2017; Karan, Sahu, and Mahto, 2020; Rahman, Lee, Islam, Antony, Ratnu, Ali, Al Mamun, Wagner, Jensen-Roberts, Waddell, Myers, Pawlik, Soto, Coffey, Sarkar, Schneider, Tarolli, Lizarraga, Adams, Little, Dorsey and Hoque, 2021). The adopted input features ranged from Mel-spectrogram and MFCC to F0, intensity, jitter, and shimmer. The classification algorithms of choice were mainly Support Vector Machines (Cortes and Vapnik, 1995) and Neural Networks (Goodfellow, Bengio, and Courville, 2016). All approaches applied a similar pre-processing: (i) extracting the feature sequences from the input speech signals, (ii) applying a pooling transformation to remove the temporal dimension from the input and obtaining a feature vector, (iii) optionally applying a dimensionality reduction transformation (e.g., Principal Component Analysis – PCA) to obtain a more compact input vector (Pearson, 1901).

Recently deep learning-based features have also been adopted for Parkison's disease detection, reporting improved performances (more than 80% recognition accuracy) with respect to solutions using standard features (Kurada and Kurada, 2020). Usually, deep learning-based features help yield better results, especially in the presence of large data sets. However, as we also show in this research, traditional prosodic and acoustic features can reach competitive performances. Moreover, traditional and handcrafted features help maintain properties like the explainability

and the interpretability of the results of the classifiers, also on problems like Parkison's disease or Alzahimer's detection (Favaro, Motley, Samus, Butala, Dehak, Oh, and Moro-Velazquez, 2022a; Favaro, Motley, Samus, Butala, Dehak, Oh, and Moro-Velazquez, 2022b).

All the existing works focused on languages individually. Most of the available results come from a benchmark data set for English (more details on this data in § 4). Lately, the attention is shifting towards multilingual solutions. However, even in cases where the classifiers were evaluated on multiple languages, none of the models trained on one language was evaluated on another to assess the cross-language performances (Toye and Kompalli, 2021). The standard approach is to train an ad-hoc model per language. Despite this being the best possible approach, it is not applicable in conditions of scarce data availability.

2.3 Domain adaptation

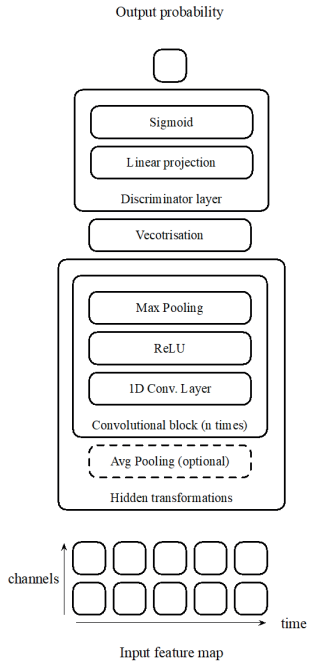
Domain adaptation is the area of machine learning that deals with the problem of applying transfer learning in cases where the original training data (the source data) are slightly different from the new inference data (the target data) (Redko, Morvant, Habrard, Sebban, and Bennani, 2019). We talk of data domain (source and target domains) to indicate the distributions from which samples come from, and of domain shift to address the change in distribution between the source and target domain. We distinguish between unsupervised and supervised domain adaptation depending whether the labels on the target data are unknown or known, respectively. In this work, we focus on unsupervised approaches even though we have target data labels to test our approach in the most complex scenario.

Unsupervised methods are designed to search for a common representation space where samples of the two domains are indistinguishable, while maintaining good performances on the original task. Through the years, many different solutions have been proposed. CORAL (Sun, Feng, and Saenko, 2016) and its deep learning variant Deep-CORAL (Sun and Saenko, 2016), focus on reducing the difference in covariance between the two domains so that the variance of the source and target distributions are indistinguishable. CORAL replaces the covariance of the source data with that of the target prior training, while Deep-CORAL adds an explicit term to the training loss of the neural network it is applied to, so that the difference in the covariance between source and target data is minimised. Gradient reversal layer (Ganin and Lempitsky, 2015) instead uses an adversarial approach and builds a classifier to distinguish between the two domains, and then inverts the gradient updates of the classifier to make the two domains undistinguishable.

Supervised (and sometimes semi-supervised) methods are easier to deal with (Redko et al., 2019). If sufficient target data points are available, it is possible to train directly on that. Otherwise, all unsupervised approaches can be adopted adding the available labels on the target data and computing the regular task loss on those labels alongside the adaptation loss. Alternatively, simple techniques like weighted losses or gradient boosting are sufficient in these settings.

3. *Neural network model*

Figure 1 -*1D convolutional neural network architecture*



In this section we describe the neural network model we used to classify audio samples into. We provide the details of the feature extraction modules to convert the raw audio into a more suitable format, the architecture of the neural network classifier and the training approach.

3.1 Input features

The proposed classifier works on 1D features maps, that are matrices of features $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ where each column-vector $\boldsymbol{x} \in \mathbb{R}^{d}$ corresponds to a specific time stamp, and each value in the vector corresponds to the value of a feature for that time stamp. Slices of features through time (rows of the input matrix) are referred to as channels. The raw input waveform is converted to this feature map, either stacking or using deep learning-based feature extraction pipelines. The model we propose is agnostic of the input features and can be trained with any of the inputs we propose hereafter.

Concerning traditional features, we followed the work of Toye and Kompalli (2021) and adopted the following prosodic and acoustic ones:

- 13 MFCC;
- F0 (pitch);
- Jitter (absolute, relative, rap, and PPQ5);
- Shimmer (absolute, absolute dB, relative, APQ3, and APQ5);
- Harmonicity.

To extract the prosodic and acoustic features, we re-sampled the audio clips at 16kHz and we computed the features in sliding windows with a width of 1024 samples and a hop of 256 samples.

For deep learning-based features, instead, we considered the three main available models for audio and speech features extraction: VGGish, SoundNet and Wav2Vec 2.0. We used all pre-trained implementations of the three neural networks. For the details about the temporal windows, we invite the readers to refer to the original papers.

3.2 Neural network classifier

The classification model $f_{PD}(\cdot)$ we are proposing is a 1D Convolutional Neural Network (CNN). The CNN, which we represented in Fig. 1, is composed by a stack of convolutional blocks, a pooling layer, and a final classification layer. Prior to the first convolutional block, we considered a layer of average pooling to reduce the temporal dimension of the input features. Note that being the CNN a fully convolutional model, it can process inputs of varying and undefined length. Hereafter, we report the general description of the architecture presented in Fig. 1; for further details on the hyperparameters of each block, refer to § 5.

The stack of convolutional blocks is used to learn a hierarchy of hidden features on top of the input ones to solve the classification problem (Parkison's disease detection). Each block is composed of three pieces: a 1D convolutional layer, a $ReLU(\cdot)$ activation and a max pooling transformation to reduce the spatial dimensionality of the input feature map. The output of one block is forwarded to the following one as input. The overall stack of transformations constitutes the hidden transformation $\boldsymbol{h}(\cdot)$ of our CNN, this function maps the input feature map $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ to another feature map $\boldsymbol{H} = \boldsymbol{h}(\boldsymbol{X}) \in \mathbb{R}^{t' \times h}$.

The pooling layer after the convolutional blocks completes the hidden transformation, deleting the temporal dimension of the feature map resulting from the hidden transformations. This step allows to compress all the information in a single vector $\boldsymbol{h} = \boldsymbol{f_{pooling}}(\boldsymbol{H}) \in \mathbb{R}^{h'}$ that can be finally used for the classification. We considered three different pooling transformations: Global Average Pooling (GAP), Global Max Pooling (GMP) and flattening (Lin, Chen and Yan, 2014). With the first two approaches we have $\boldsymbol{h'} = \boldsymbol{h}$, while with the latter $\boldsymbol{h'} = \boldsymbol{t'} \cdot \boldsymbol{h}$.

The final classification layer is composed by a learnable linear projection that maps the hidden vector $\boldsymbol{h}$ to a scalar value, the logit. A final sigmoid activation function $\sigma(\cdot)$ allows to output the probability of the input sample speech being from a subject affected by Parkinson's disease.

3.3 Training approach

The loss function $\mathcal{L}(\cdot)$ optimised to train the neural network is a linear combination of two losses: the former is the usual binary cross-entropy (BCE) loss $\mathcal{L}_{BCE}(\cdot)$, and the latter is the Deep-CORAL objective $\mathcal{L}_{DC}(\cdot)$ to achieve unsupervised domain adaptation. Given the labelled source data $D_s = \{(\boldsymbol{X_{S,1}}, y_{S,1}), \ldots, (\boldsymbol{X_{S,n}}, y_{S,n})\}$,

composed of input-target output pairs (with $y_{S,i} \in \{\mathbf{0}, \mathbf{1}\}\ \forall \boldsymbol{i}$) and the unlabelled target data $D_T = \{\boldsymbol{X_{T,1}}, \ldots, \boldsymbol{X_{T,m}}\}$

$$\mathcal{L}(D_s, D_T) = \mathbb{E}_{(\boldsymbol{X_{S,I}}, y_{S,i}) \in D_s}[\mathcal{L}_{BCE}(\boldsymbol{X_{S,I}}, y_{S,i})] + \lambda \mathcal{L}_{DC}(D_s, D_T)$$

where $\lambda \in [0, 1] \subseteq \mathbb{R}$ is a parameter to control the contribution of the adaptation loss (note that $\lambda = 0$ disables the adaptation).

The binary cross-entropy loss $\mathcal{L}_{BCE}(\cdot)$ is computed as

$$\mathcal{L}_{BCE}(\boldsymbol{X_{S,i}}, y_{S,i}) = y_{S,i} \cdot \log(\boldsymbol{f_{PD}}(\boldsymbol{X_{S,i}})) + (\mathbf{1} - y_{S,i}) \cdot \log(\mathbf{1} - \boldsymbol{f_{PD}}(\boldsymbol{X_{S,i}}))$$

The Deep-CORAL loss for adaptation is computed as

$$\mathcal{L}_{DC}(D_s, D_T) = \frac{1}{4h'^2} \| \Sigma_S - \Sigma_T \|_F^2$$

where $\Sigma_S \in \mathbb{R}^{h' \times h'}$ and $\Sigma_T \in \mathbb{R}^{h' \times h'}$ are the covariance matrices of the hidden feature vectors computed from the samples in the source and target data sets respectively, and $\|\cdot\|_F^2$ is the squared Frobenius matrix norm operator.

4. *Data*

In this section, we present the audio samples we used in this research and all the pre-processing and preparation steps we applied to the available samples.

4.1 Data set

For the scope of this work, we considered two distinct data sets composed by speech samples coming from persons either affected by Parkinson's disease or in healthy conditions. The former data set contains speech recordings on English speakers, while the latter contains speech recordings of Telugu speakers.

The English data set is a known data set for Parkinson's disease detection: the Mobile Device Voice Recordings at King's College London (MDVR-KCL) (Jaeger, Trivedi, and Stadtschnitzer, 2019). The recordings are phone calls sampled at 44.1 kHz 16 bit. Samples come from patients affected by Parkison's disease at different stages of the disease and from patients in healthy conditions. We have samples of female speakers and male speakers from different age ranges (age distribution of patients affected by Parkinson's disease is higher to the disease affecting mostly elderly people). Of the available recordings, 35 m 38 s (corresponding to 31 samples) come from patients affected by Parkinson's disease and 46 m 5 s (corresponding to 42 samples come from patients in healthy conditions.

The Telugu data set is the motivation behind this work. We were provided with a private collection of 12 m 39 s of audio recordings from native Telugu speakers affected by Parkinson's diseases, for a total of 57 samples. Since all the samples were coming from unhealthy patients, we sampled randomly 200 sample (15 m 35 s of recordings) of Telugu speakers in healthy conditions from the Telugu split of the Open Speech and Language Resources (OpenSLR) corpus (He, Chu, Kjartansson,

Rivera, Katanova, Gutkin, Demirsahin, Johny, Jansche, Sarin, and Pipatsrisawat, 2020), a data set of audio books for speech recognition, to obtain the healthy patients' samples. We took half of the samples from male speakers and half from female speakers. We do not have access to the gender information of private collection, and we do not have any information on age distribution of both data sets.

4.2 Pre-processing and preparation

Figure 2 - *Audio samples duration distributions divided per language (the green line is the combine distribution of both kind of patients, values refer to duration after sentence boundaries cropping)*

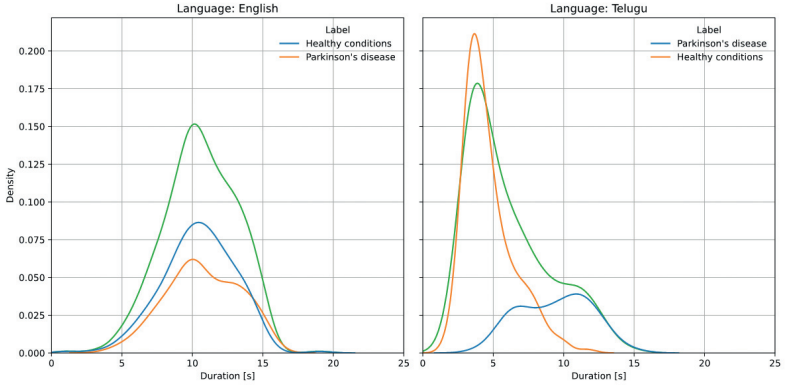


Table 1 - *Audio samples statistics organised per language and per label (PD: Parkinson's Disease, HC: Healthy Conditions) and divided into training and test split, counts are taken after sentence boundaries splitting and after windowing (total counts aggregate across data splits, both counts aggregate across labels)*

Language	*Label*	*Training*		*Test*		*Total*	
		Sentences	*Windowed*	*Sentences*	*Windowed*	*Sentences*	*Windowed*
English	*PD*	148	465	51	149	199	614
	HC	197	591	64	193	261	784
	Both	345	1056	115	352	460	1480
Telugu	*PD*	57	153	24	70	81	223
	HC	153	244	47	71	200	315
	Both	210	397	71	141	281	538

Given the available samples, we applied the same pre-processing steps to both data sets. First, we applied denoising to make sure that only human voice would be present in the recording, then we cropped the audio clips in correspondence of sentence boundaries. Finally, we applied windowing to the resulting samples sample. The neural network classifies individual windows.

To denoise the audio clip we used the RNNoise tool from Mozilla (Valin, 2018). The tool uses a neural network to estimate the spectral components of noise in

each audio recording and applies spectral subtraction to get rid of the noise, while preserving human voice in the recording as clear as possible.

Sentence boundaries splitting was applied manually on both data sets. We looked for long pauses or points where we were sure there was some syntactic boundary, and we divided the original audio sample. We did not apply this process to the OpenSLR part of the Telugu data. The resulting audio samples duration distribution and the actual number of samples resulting from this process are reported, respectively, in Fig. 2 and Tab. 1.

We applied a further windowing to the resulting samples. We cropped the recordings in windows of 4 s with a stride of 2 s between one window and the other. The resulting number of windowed samples is reported in Fig. 2 as well.

5. *Experiments*

We experimented training different configurations of the neural network (changing the input features and the pooling approach) and we compared results on source language as well as results on target language. For the latter we considered the results with and without adaptation. The main experiment was on the adaptation from English to Telugu; however, we also evaluated the other way around for completeness. In both cases, we applied a train-test split on the source language data (75%-25% split) and used the entire target data set for evaluation.

To complete data preparation, we applied standardisation to the input features to have all the feature vectors in the input feature maps have zero mean and unit standard deviation. We estimated separately the statistics for source data (on the training set) and target data (on the entire data set). Additionally, we applied minority oversampling to the source training data to have a balanced training set.

We considered the four different input features introduced in § 3.1 and the three different pooling approaches introduced in § 3.2. For each feature and pooler combination we run a 5-fold cross-validation to find the best configuration. For every fold, we reserved 20% of the training data for validation and early stopping (to prevent overfitting). After finding the best configuration we re-trained the best configuration using all the available source training data (maintain the 20% split for validation).

As training hyperparameters we set the batch size to 16, the maximum number of training epochs to 150 and the early stopping patience to 5. For the neural network we used separable 1D convolution with the kernel size set to 3 and the stride on 1, while for max pooling we set the kernel size to 2 and the stride to 2. Before each convolution, we applied dropout with a rate of 10% for regularisation. We trained all configurations using ADAM as optimiser (Kingma and Ba, 2015). During the cross-validation we tuned the following hyperparameters:

- Number of convolutional blocks: 1, 2, or 3.
- Number of hidden units in the convolutions: 512 or 1024 (we used the same value for all convolutional blocks).
- Learning rate: 10^{-3} or $5 \cdot 10^{-4}$.

Due to the high density in the temporal dimension of traditional features and Wav2Vec 2.0 features, we used the optional initial average pooling layer with a kernel size of 64 elements and a stride of 32.

Concerning adaptation, we used the best configuration of each input feature-pooling approach pair found on the source data also for the adaptation experiment. During training with adaptation, we set the weight of the Deep-CORAL loss to $\lambda = 1/10$.

For evaluation, we computed the usual classification metrics: accuracy, precision, recall, F1, specificity (negative class recall), and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. For the target language, we compared the results with and without the adaptation, to better understand its effects.

6. *Results*

Figure 3 - *English Parkinson's disease detection classifiers results*

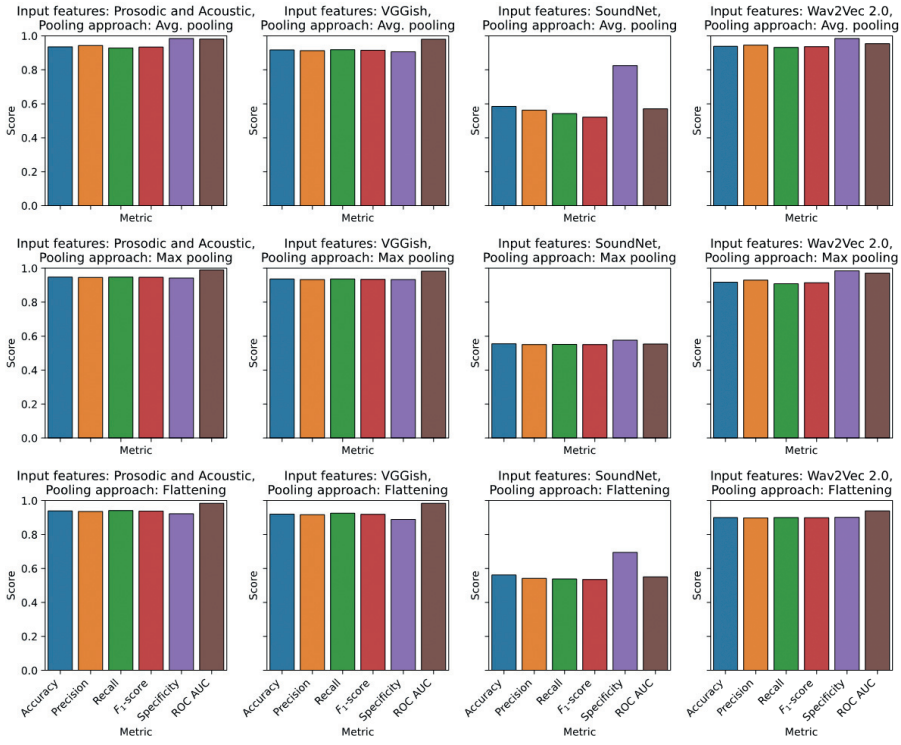
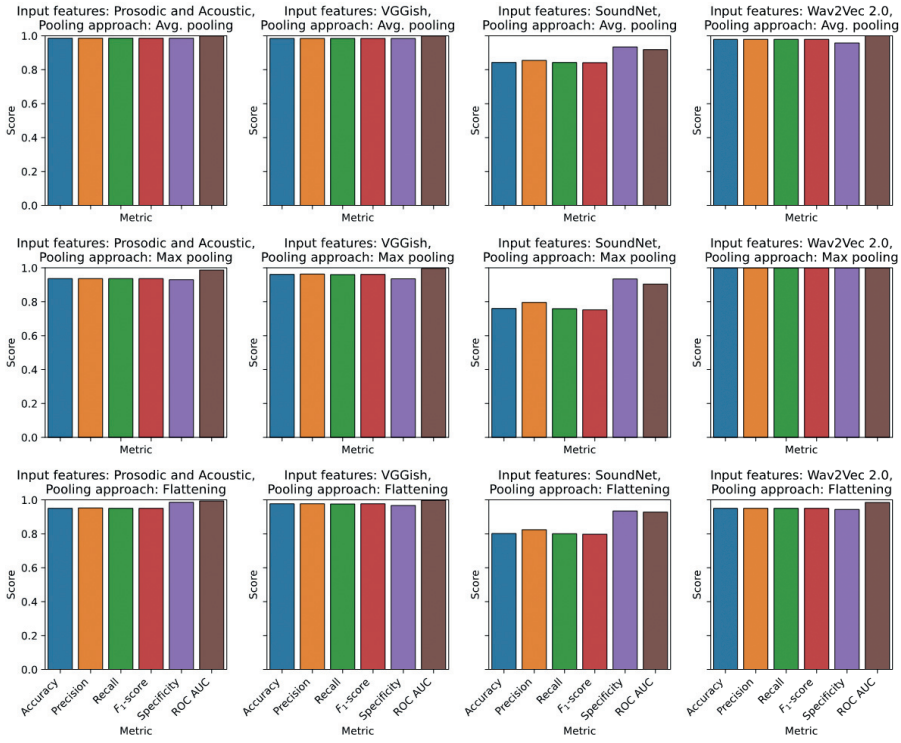


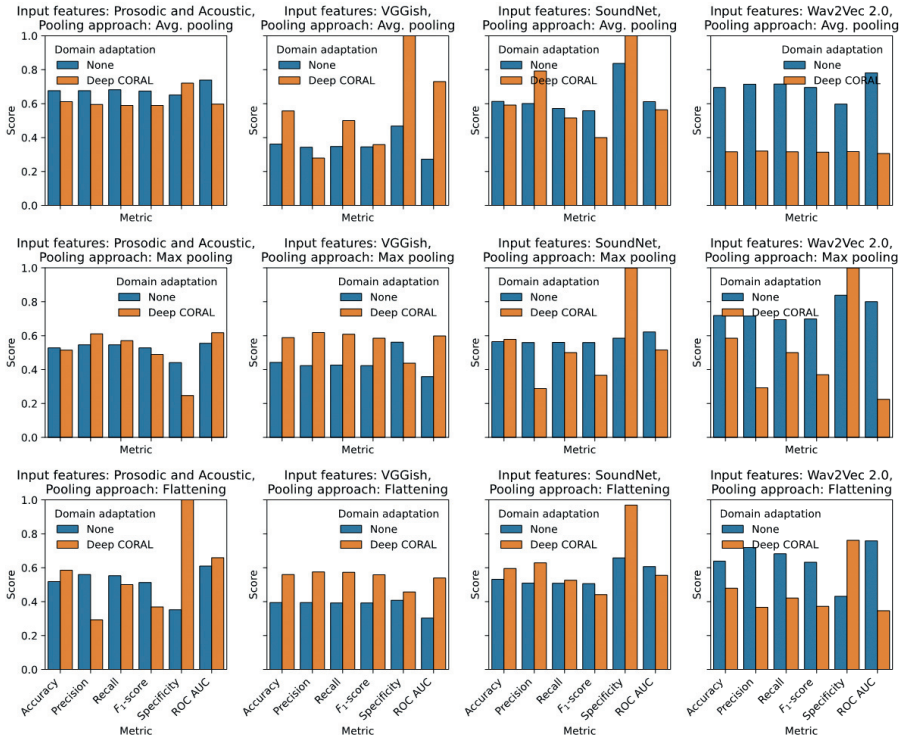
Figure 4 - *Telugu Parkinson's disease detection classifiers results*



We reported the results on the test split of the source languages in Fig. 3 (for English) and Fig. 4 (for Telugu). These results were obtained separately after training on the test split for that same source language.

Results are similar between English and Telugu, although having Telugu fewer samples, they are less robust. On both languages we managed to obtain impressive recognition results. Scores are >90% on all metrics with all input features apart from SoundNet. Traditional features achieve competitive results at the same level of VGGish and Wav2Vec 2.0. About SoundNet, we hypothesised that the lower scores are due to the feature extractor not having "seen" human voice in the pre-training.

Figure 5 - *Comparison of results of domain adaption using Deep-CORAL, on Parkinson's disease detection from English (source language) to Telugu (target language), with respect to no-adaptation baseline*



Concerning adaptation from English to Telugu, we reported the results with and without Deep-CORAL loss in Fig. 5. In general, we obtained mixed results. On traditional features the results with and without adaptation are mostly undistinguishable. Same applies for SoundNet; however, from the source data results analysis, SoundNet features seemed to be useful to this classification task. On VGGish we noticed the expected behaviour, results without adaption were mostly below 50% and Deep-CORAL helped improve the performances on the target language. The result is not as sharp as we expected, but given the unsupervised approach and the reduced data set, it is still valuable. Finally, from Wav2Vec 2.0 we notice the opposite behaviour. The model performed well out-of-the-box with the target language, and the adaptation decreased the performances. We hypothesise that the cross-language results without adaptation are to the self-supervised pre-training of Wav2Vec 2.0 feature extractor, that was conducted on a large language pool (Baevski et al., 2020).

Adaptation from Telugu to English did not highlight any significant result. However, given the reduced size of the source data set, we were not expecting any result in this sense and we run the experiments only for completeness.

7. *Conclusions and future works*

In this paper we explored domain adaptation applications to Parkinson's disease detection, adapting from a source language to a target one. The scope of the research was to understand if, in presence of reduced and unlabelled data sets it would be possible to resort to this kind of techniques. We selected English as source data and Telugu as target data, and we explored different input features to feed a CNN, both traditional prosodic and acoustic features and newer deep learning-based features.

From the results we collected, we noticed that some deep learning-based features obtained via self-supervised pre-training on multiple-languages (Wav2Vec 2.0) are naturally compatible with languages unseen during the transfer learning and are actually badly affected from domain adaptation (results were better when applying an English classifier directly to Telugu rather than using adaptation). However, in the cases where the performances without adaptation were low on the target data, the selected domain adaptation technique correctly helped improve the results (e.g., the case of VGGish based classifiers). Given that we were approaching the problem in a completely unsupervised way, we expect that having access to a larger target data set (even if unlabelled) would help improve the results.

The next steps on this work are to scale up the experiments. We are interested in considering more languages and more adaptation techniques. In fact, we are willing to see whether we can directly train a multilingual model. Moreover, we want to understand if other adaptation techniques can help yield better results.

Acknowledgements

The authors wish to thank Prof. Anitha S. Pillai of Hindustan Institute of Technology and Science for providing the audio samples of Parkinson's disease patients in Telugu.

References

Aytar, Y., Vondrick, C. & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon & Roman Garnett (a cura di), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, Barcelona, Spain, 5-10 December 2016, 892–900. URL https://proceedings.neurips.cc/paper/2016/hash/7dcd340d84f762eba80aa538b0c527f7-Abstract.html

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan & Hsuan-Tien Lin (a cura di), *Advances in Neural Information Processing Systems: 354th Annual Conference on Neural Information Processing Systems* (NeurIPS), December 6-12 2020, virtual, vol. 33, 12449-12460 URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020*). Language models are few-shot learners.* In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (a cura di), *Advances in Neural Information Processing Systems: 354th Annual Conference on Neural Information Processing Systems* (NeurIPS), December 6-12 2020, virtual, vol. 33, 1877-1901, URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac 142f64a-Abstract.html

Chernykh, V., Sterling, G. & Prihodko, P. (2017). *Emotion recognition from speech with recurrent neural networks*. In ArXiv abs/1701.08071. URL http://arxiv.org/abs/1701.08071

Cortes, C. & Vapnik, V. (1995). Support-vector networks. In *Mach. Learn*., 20(3):273–297, doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018

Dimauro, G., Caivano, D., Bevilacqua, V., Girardi, F. & Napoletano, V. (2016). Voxtester, software for digital evaluation of speech changes in parkinson disease. In *IEEE International Symposium on Medical Measurements and Applications, MeMeA,* Benevento, Italy, 15-18 May 2016, 1–6, doi: 10.1109/MeMeA.2016.7533761. URL https://doi.org/10.1109/MeMeA.2016.7533761

Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D. & Girardi, F. (2017). Assessment of speech intelligibility in parkinson's disease using a speech-to-text system In *IEEE Access*, vol. 5, 22199–22208, doi: 10.1109/ACCESS.2017.2762475. URL https://doi.org/10.1109/ACCESS.2017.2762475

Favaro, A., Motley, S., Samus, Q.M., Butala, A., Dehak, N., Oh, E.S. & Moro-Velazquez, L. (2022). A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. In IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar,532-539, doi: 10.1109/SLT54892.2023.10022435

Favaro, A., Motley, S., Samus, Q.M., Butala, A., Dehak, N., Oh, E.S. & Moro-Velazquez, L. (2022). Artificial intelligence tools to evaluate language and speech patterns in alzheimer's disease. In *Alzheimer's & Dementia*, 18(S2):e064913, doi: https://doi.org/10.1002/alz.064913. URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.064913

Giannakopoulos, T. (2015). Pyaudioanalysis: An open-source python library for audio signal analysis. In *PLOS ONE*, 10(12):1–17, doi: 10.1371/journal.pone.0144610. URL https://doi.org/10.1371/journal.pone.0144610

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Francis R. Bach & David M. Blei, (a cura di), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, Lille, France, 6-11 July 2015, vol. 37, 1180–1189. URL http://proceedings.mlr.press/v37/ganin15.html

Gao, J., Galley, M. & Li, L. (2018). Neural approaches to conversational AI. In Artzi, Y. & Eisenstein, J. (a cura di), *Proceedings of ACL 2018,* Melbourne, Australia, 15-2 July 2018, Tutorial Abstracts, 2–7. doi: 10.18653/v1/P18-5002. URL https://aclanthology.org/P18-5002/

Goodfellow, I.J., Bengio, Y. & Courville, A.C. (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press, ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/

He, F., Cathy Chu, S., Kjartansson, O., Rivera, C., Katanova, A., Gutkin, A., Demirsahin, I., Johny, C., Jansche, M., Sarin, S. & Pipatsrisawat. K. (2020). Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC), Marseille, France, 20-25 June, 6494–6503. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.800

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Channing Moore, R., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J. & Wilson, K.W. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), New Orleans, LA, USA, 5-9 March 2017, 131–135, doi: 10.1109/ICASSP.2017.7952132. URL https://doi.org/10.1109/ICASSP.2017.7952132

Jaeger, H., Trivedi, D. & Stadtschnitzer, M. (2019). *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls*, Zenodo, May 2019, URL https://doi.org/10.5281/zenodo.2867216

Karan, B., Sekhar Sahu, S. & Mahto, K. (2020). Parkinson disease prediction using intrinsic mode function based features from speech signal. In *Biocybernetics and Biomedical Engineering*, 40(1):249–264.

Kingma D.P. & Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (a cura di). In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 7-9 May 2015, URL http://arxiv.org/abs/1412.6980

Kurada, S. & Kurada. A. (2020). Vggish embeddings based audio classifiers to improve parkinson's disease diagnosis. In *5th IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (*CHASE 2020), Crystal City, VA, USA, 16-18 December, 9–11, doi: 10.1145/3384420.3431775. URL https://doi.org/10.1145/3384420.3431775

Lin, M., Chen, Q. & Yan, S. (2014). *Network in network.* In Bengio, Y. & LeCun, Y. (a cura di), *2nd International Conference on Learning Representations* (ICLR 2014), Banff, AB, Canada, 14-16 April, 2014. URL http://arxiv.org/abs/1312.4400

Merler, M., Joshi, D., Mac, K.C., Nguyen, Q., Hammer, S., Kent, J., Xiong, J., Do, M.N., Smith, J.R. & Schmidt Feris, R. (2018). The excitement of sports: Automatic highlights using audio/visual cues. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops),* Salt Lake City, UT, USA, 18-22 June 2018, 2520–2523. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w49/html/Merler_The_Excitement_of_CVPR_2018_paper.html

Pearson, K. (1901). Liii. On lines and planes of closest fit to systems of points in space. In *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P.J. (2019) Exploring the limits of transfer learning with a unified text-to-

text transformer. In *Journal of Machine Learning Research*, vol. 21(1:140), 5485-5551. URL http://jmlr.org/papers/v21/20-074.html

Rahman, W., Lee, S., Islam, M.S., Antony, V.N., Ratnu, H., Ali, M.R., Mamun, A.A., Wagner, E., Jensen-Roberts, S., Waddell, E., Myers, T., Pawlik, M., Soto, J., Coffey, M., Sarkar, A., Schneider, R., Tarolli, C., Lizarraga, K., Adams, J., Little, M.A., Dorsey, E.R. & Hoque, E. (2021). Detecting parkinson disease using a web-based speech task: Observational study. *Journal of medical Internet research*, vol. 23(10):e26305, 2021.

Redko, I., Morvant, E., Habrard, A., Sebban, M. & Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.. URL https://www.elsevier.com/books/advances-in-domain-adaptation-theory/redko/978-1-78548-236-6

Schneider, S., Baevski, A., Collobert, R. & Auli, M. (2019). *wav2vec: Unsupervised pre-training for speech recognition*. In Gernot Kubin & Zdravko Kacic (a cura di). In *Proceedings of Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, 15-19 September 2019, 3465–3469. doi: 10.21437/Interspeech.2019-1873. URL https://doi.org/10.21437/Interspeech.2019-1873

Scotti, V., Galati, F., Sbattella, L. & Tedesco, R. (2020). Combining deep and unsupervised features for multilingual speech emotion recognition. In Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J. & Vezzani, R. (a cura di). *In Pattern Recognition, ICPR International Workshops and Challenges* – Virtual Event, January 10-15, 2021, Proceedings, Part II, vol.12662 of Lecture Notes in Computer Science, 114–128, Springer. doi: 10.1007/978-3-030-68790-8_10. URL https://doi.org/10.1007/978-3-030-68790-8_10

Sun, B. & Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. In Hua, G. & Jégou, H. (a cura di). In *Proceedings of the Computer Vision – ECCV 2016 Workshops*, Amsterdam, The Netherlands, 8-10 October 8-10 2016, Part III, vol. 9915 of Lecture Notes in Computer Science, 443–450. doi: 10.1007/978-3-319-49409-8_35. URL https://doi.org/10.1007/978-3-319-49409-8_35

Sun, B., Feng, J. & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In Schuurmans, D. & Wellman, M.P. (a cura di). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 12-17 February 2016, Phoenix, Arizona, USA, 2058–2065. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12443

Tan, X., Qin, T., Soong, F.K. & Liu, T. (2021). A survey on neural speech synthesis. In *CoRR*, abs/2106.15561. URL https://arxiv.org/abs/2106.15561.

Toye, A.A. & Kompalli, S. (2021). Comparative study of speech analysis methods to predict Parkinson's disease. In CoRR, abs/2111.10207. URL https://arxiv.org/abs/2111.10207

Valin, J. (2018). A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *20th IEEE International Workshop on Multimedia Signal Processing (MMSP 2018),* Vancouver, BC, Canada, 29-31 August 2018, 1–5. IEEE, 2018. doi: 10.1109/MMSP.2018.8547084. URL https://doi.org/10.1109/MMSP.2018.8547084

Williamson, J.R., Quatieri, T.F., Helfer, B.S., Perricone, J., Ghosh, S.S., Ciccarelli, G.A. & Mehta, D.D. (2015). Segment-dependent dynamics in predicting parkinson's disease. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 6-10 September, 2015, 518–522. URL http://www.isca-speech.org/archive/interspeech_2015/i15_0518.html

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks?. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., D. & Weinberger, K.Q. (a cura di). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, 8-13 December 2014, 3320–3328. URL https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html

Appendix

For transparency and reproducibility, we release the developed source code via GitHub at the following link: https://github.com/vincenzo-scotti/ voice_analysis_parkinson