An Analysis of Features for Machine Learning Approaches to Parkinson's Disease Detection

Claudio Ferrante[1], Bindu Menon[2], Anitha S. Pillai[3],
Licia Sbattella[4] and Vincenzo Scotti[5]

[4]https://orcid.org/0000-0001-5344-5976

[5]https://orcid.org/0000-0002-8765-604X

[1,4,5]DEIB, Politecnico di Milano
Via Golgi 42, 20133, Milano (MI), Italy

[2]Apollo Speciality Hospitals
5/639, Rajiv Gandhi Salai, Tirumalai Nagar, Perungudi, Chennai, Tamil
Nadu 600096, India

[3]Hindustan Institute of Technology and Science,
Rajiv Gandhi Salai (OMR), Padur, Kelambakkam, Tamil Nadu 603103,
India
[1]claudio.ferrante@mail.polimi.it, [2]bindu.epilepsycare@gmail.com,
[3]anithasp@hindustanuniv.ac.in, [4]licia.sbattella@polimi.it,
[5]vincenzo.scotti@polimi.it

Abstract

Natural Language Processing has been transformed by the introduction of *deep learning models* for text and speech processing. Through intuitive techniques like *transfer learning* and *fine-tuning*, it is possible to train impressive discriminative and generative probabilistic models. Moreover, these approaches seem to be able to cope with an issue that traditional handcrafted features sometimes fail to handle: lack of resources like data or domain-specific knowledge. To test the applicability of these features, we build a classification pipeline for Parkinson's disease from speech, using samples in Telugu. The data set we use is relatively small, especially if compared with those available for English for example; moreover, Hindi does not offer access to many models specific for its processing, again differently from highly resourced languages. In these settings, we evaluate the generalization capabilities of classification models using deep-learning features. In our experiments, we compare different deep-learning models and traditional prosodic and acoustic features to understand which features are the most suitable for this situation. The results are mixed: although in line with the state-of-art, the best scores we obtained are with one of the models using deep learning features, we still managed to achieve impressive scores using handcrafted features.

Keywords: Natural Language Processing, Deep-Learning, Speech analysis, Parkinson’s disease.

1. Introduction

Deep learning has slowly become an essential tool for Natural Language Processing. Apart from the impressive results on text processing [1] [2], deep learning models allowed to improve the results of multiple speech-related applications, like Automatic Speech Recognition (ASR) [3] [4], speaker identification [5], conditioned Text-to-Speech synthesis [6] [7] [8] or speech emotion recognition [9].

These deep learning models are particularly useful in the presence of small data sets or under-resourced problems (in terms of data and domain-knowledge availability). They allow to exploit techniques like transfer learning and fine-tuning [10], where the features computed by a deep learning neural network model trained on a large generic data set are re-used on a specific problem with a smaller data set, generally resulting in improved performances.

In this work, we focus on Parkinson's disease detection from speech. We propose a probabilistic classification pipeline to detect if a patient is affected by Parkinson's disease by analyzing voice recordings. To evaluate the generalization capabilities enabled by deep learning models for audio/speech feature extraction, we test them on a relatively small data set of samples in Telugu. These settings represent a challenge due to the reduced data set size and the scarce availability of language-specific analysis models.

We divide this chapter into the following sections. In Section 2 we present the related works in terms of features for speech analysis and results on Parkinson's disease detection from voice. In Section 3 we present the abstract classification pipeline we adopted, suggesting possible implementations of the various modules. In Section 4 we describe the data set we collected to train and evaluate different classification models. In Section 5 we describe the implemented pipeline configurations we evaluated and we report the results obtained during the evaluation. Finally, in Section 6 we summarize our work and suggest possible future evolution.

2. Related Works

In this section, we present the features commonly employed for speech classification, both traditional and deep learning based.

Additionally, we present the latest results for Parkinson's disease detection from speech.

2.1. Features for speech analysis

Traditionally, the speech features adopted for NLP are divided into two groups: prosodic features and acoustic features [11] [12]. The former group includes features used to describe peculiarities of speech, like: Pitch, Intensity, Harmonicity, Jitter, Shimmer, Speech Rate, Short-Term Energy, Short-Term Entropy, etc. The latter group includes the features used to describe the acoustic properties of speech, like Spectrogram (magnitude or power), Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCC), Spectrogram statistics (centroid, spread, flux, rolloff, entropy), Chromagram, etc.

More recent approaches, instead, propose to re-use deep learning models trained on large data collections. The internal representations learned by these models are particularly informative and can be directly transferred or easily adapted to many new problems. The most popular models in this sense are SoundNet [13], VGGish [14], and Wav2Vec [15] [3]. The first two are very generic models, though for acoustic analysis not necessarily aimed at speech. SoundNet is a 1D convolutional neural network trained to predict from the audio track of video clips the pseudo labels generated from an object recognition deep neural network and a scene recognition deep neural network that processed the images of the video clips. VGGish, on the other hand, is a 2D convolutional neural network trained on a large audio classification data set containing a large number of labels and samples. Instead, both the versions of Wav2vec were specifically designed for speech analysis problems and were originally used as input for state-of-the-art ASR models.

2.2. Parkinson’s disease detection from speech

Parkinson's disease detection from speech has already been explored as a machine-learning problem.

More sophisticated solutions also adopted dimensionality reduction techniques to feed more compact and informative feature vectors (encoding the input speech signal) to the classifiers. Different classification algorithms, like Artificial Neural Networks, Support Vector Machines, and k-Nearest Neighbors, have been adopted for this detection problem [16] [17] [18] [19] [20] [21] [22]. Usually, these solutions involved the extraction of prosodic and acoustic features (mainly MFCC, Jitter, Shimmer, and Pitch), which allowed to train discriminative models with impressive results. In some cases, these features were further processed through dimensionality reduction transformations to keep only the most relevant components of the vectors encoding the audio clips to classify, which resulted in further improvements is some cases.

Recent results also explored the effect of deep learning features to train a classifier for this Parkinson's disease detection problem [19], reaching more than 80% recognition accuracy on a data set of English audio clips, outperforming the other considered classifiers based on spectral features. Other works started focusing on building classifiers compatible with multiple languages [20]: the results showed that acoustic and spectral features can be used to build high-performing classifiers (reaching more than 90% recognition accuracy) on English and Italian.

All these works relied on larger data sets like Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls [23] and Italian Parkinson's Voice and Speech [24] [25], which account for more than 1 h of recordings. In our case we are dealing with a much smaller data set, thus we are interested in seeing if and how much performance degrades when using similar classification pipelines.

3. Proposed machine learning pipeline

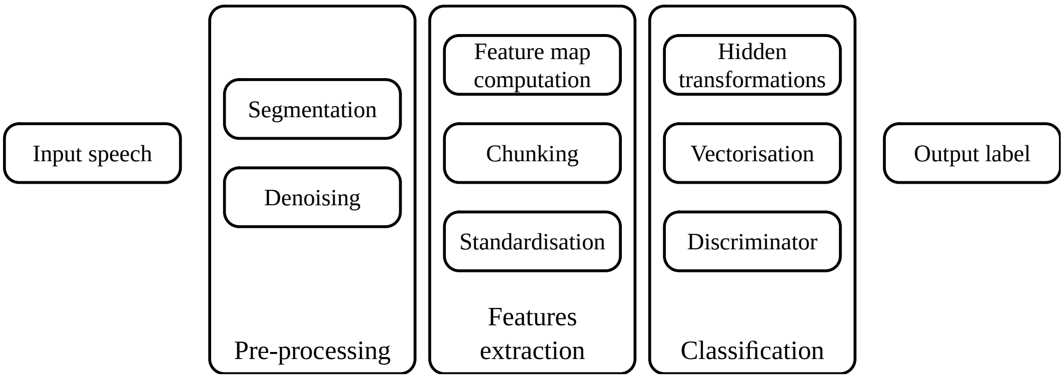

Figure 3.1. Visualization of the abstract classification pipeline.

In this section, we describe the classification pipeline we propose for Parkinson's disease detection from speech; we depicted the pipeline in Figure 3.1. We compose each of the stages of the pipeline (preprocessing, feature extraction and classification) of different modules. The choice of specific module implementations allows to instance the proposed pipeline into a classification model, which can be trained and evaluated.

3.1. Pre-processing

In our pipeline, we considered two pre-processing steps: segmentation and denoising. Both steps prepare the raw data for the feature extraction stage.

Segmentation consists of the splitting of the audio clip in presence of longer pauses, which generally mark the end of an utterance. This step can be done by hand (however it may require a lot of time) or automatically. In the latter case, it is better to do it after denoising, to avoid errors due to additional sounds present in the recording that overlap with the voice.

The denoising module takes care of removing (as much as possible) additional signals in the input audio clip which overlap with the voice to analyze. State-of-the-art solutions use deep learning models trained on many hours of data and can achieve impressive results. This is an important step; due to the reduced size of the data set we are working with; we cannot expect the classification models to generalize and learn implicitly to ignore the noise.

3.2. Features extraction

Features extraction is the core stage of our pipeline. The main step at this stage is the computation of the feature map, which, after the appropriate post-processing steps is the actual input of the classification model.

Each pre-processed input speech signal undergoes a transformation to extract a feature map. Traditionally data samples for machine learning are represented by a $d \in \mathbb{N}$ feature vector. However, for some problems like speech or image analysis, it is possible to leverage the spatial structure of the input and generate a feature map. In our case, from a speech signal, we can compute a feature map that is a sequence of feature vectors (each computed is a specific time window of the input signal) that can be encoded in a matrix $\boldsymbol{X} \in \mathbb{R}^{t \times d}$, where t is the number of time positions and d is the number of features for each vector. Deep learning models, as well as algorithms to compute traditional prosodic and acoustic features, iteratively transform the raw input signal to obtain a feature map.

Depending on the duration of the input audio clip, we considered the possibility of an intermediate chunking module. To avoid processing too large segments of audio, which can be computationally expensive and may harm the results in the presence of shorter input, we introduced an optional chunking step. The feature maps can be chunked into smaller windows along the time axis, to process smaller portions of the input speech. In this way, from the same segment, it is possible to extract multiple samples for the classifier.

The last step in the features extraction stage is standardization. This transformation is used to mean-center the data and impose a variance of 1 for all the features individually. Standardization is used to have the same scale on all the features, which helps the convergence and stability of the learning algorithm used for classification. There exists some robust version of this transformation using, for example, the median instead of the mean for centering.

3.3. Classification

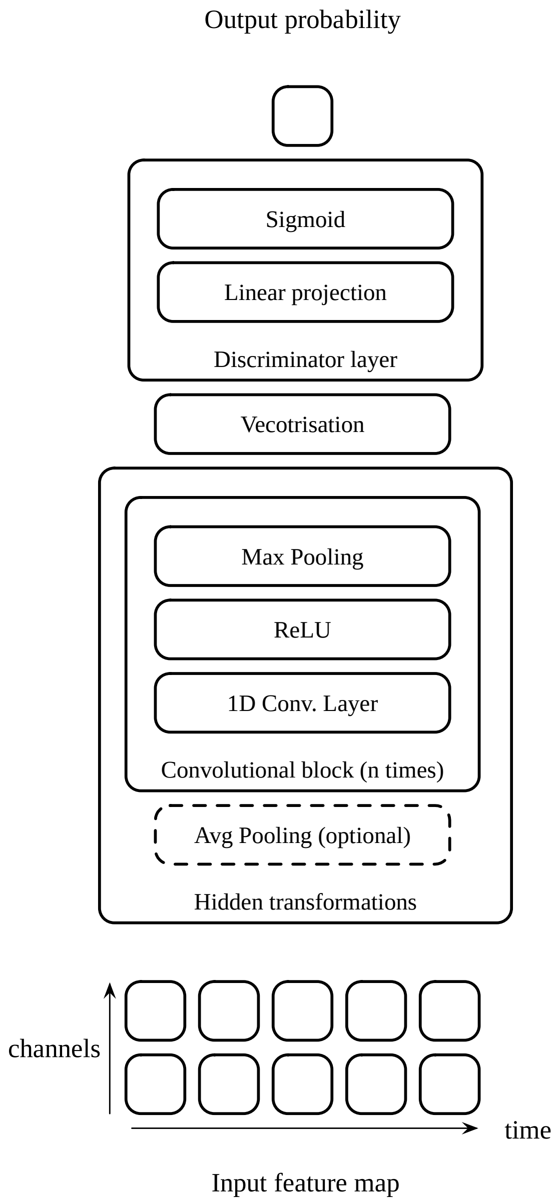

Figure 3.2. Convolutional Neural Network Classifier Architecture.

The last step of the proposed pipeline is a Convolutional Neural Network (CNN) classifier [26]. We approach the problem as a supervised binary classification problem. In input we have a feature map extracted from an audio clip of human speech and in output we have the probability for that clip to correspond to a Parkinson's Disease patient. We report a diagram of the considered architecture in Figure 3.2.

The CNN is composed of a hidden transformation $h(\cdot)$, composed of convolutional blocks. These transformations process the input feature map $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ transforming it into another feature map $\boldsymbol{H} \in \mathbb{R}^{t \times h}$ to be used for the final classification. The convolutional blocks contain, in order, a 1D convolutional layer, a $ReLU(\cdot)$ activation and a max pooling transformation to reduce the spatial dimensionality of the feature map. Optionally, we considered an initial average pooling transformation to reduce the spatial dimensionality before the convolutional blocks. This optional transformation is useful when dealing with highly dense feature maps like those from handcrafted features or from Wav2Vec 2.0.

After the hidden transformation, we apply a vectorization transformation. The role of this transformation is to drop the spatial dimension and convert the feature map $\boldsymbol{H} \in \mathbb{R}^{t \times h}$ into a feature vector $\boldsymbol{h} \in \mathbb{R}^{h}$ to be classified. We considered the following vectorization approaches: flattening (or unrolling), Global Average Pooling (GAP) [30], and Global Max Pooling (GMP) [30].

The last layer of the CNN is a linear transformation that maps the feature vector $\boldsymbol{h} \in \mathbb{R}^{h}$ into a scalar value. This value is passed through a Sigmoid activation function to have the output probability score.

4. Data

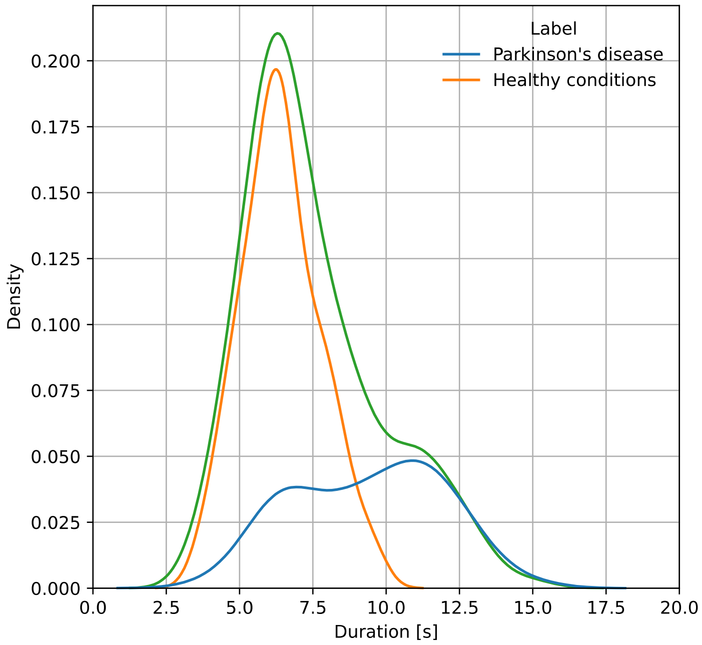

Figure 4.1. Distribution of audio clips duration after segmentation. The green line is the overall distribution including all samples.

As discussed, in this work we adopted a data set of audio clips collected from Telugu speakers. The reasons behind this choice are two: there, thus we wanted to establish a baseline, and we wanted to see if deep learning features for speech analysis allow generalizing on under-resourced languages and small data sets, like in this case.

The data set we adopted is composed of two parts. The samples from Parkinson's disease patients come from a private data set, composed of 12 m 39 s of audio recordings. To balance this data set with samples from healthy persons (in the sense of speakers not affected by Parkinson's disease), we gathered the audio samples from the delta segment of the Telugu split of the Open Speech and Language Resources (OpenSLR) corpus [28], which accounts for 15 m 35 s of recordings. The total amount of available data is 28 m 14 s. In terms of samples, we collected a total 281 audio clips, 200 from patients in healthy conditions and 71 from patients affected by Parkinson's Disease.

Before processing the audio clips with the feature extraction modules of our pipeline, we manually segmented those clips. We split the recordings on longer pauses, which we associated with utterances boundaries. In Figure 4.1 we displayed the distribution of the audio clip lengths after the manual segmentation. As can be seen, audio clips of Parkinson's disease patients cover a wider duration range than those from healthy patients. This difference is primarily due to the different sources of the data samples, other than the articulation difficulties due to the disease.

As we explained in Section 3, the speech samples are analyzed in smaller chunks, to prevent overfitting of the classifiers, given the reduced size of the data set.

5. Evaluation and results

In this section, we describe the experiments we conducted on different models and how we evaluated their performances. Additionally, we present and comment on the obtained results.

5.1. Experiments

Our experiments were mainly though to compare different modules for the feature extraction stage, in order to identify the most suitable features for Parkinson's disease from speech on the considered Telugu data set. In this sense, we compared different feature map computation approaches, analyzing the results obtained using features from different deep learning models. Additionally, we explored different vectorization algorithms, applying them to the extracted feature maps.

In the preprocessing stage, after the manual segmentation step to isolate the different utterances in the same recording, we used a denoising application to enhance voice and remove background noises that could have affected the classifier. We employed RNNoise [29] tool to denoise the input speech segments.

For the feature extraction stage, we compared three different pre-trained deep learning models for acoustic features extraction: SoundNet, VGGish, and Wav2Vec 2.0. To have a term of comparison with these deep learning features, we also trained some models using traditional (handcrafted) speech analysis features. Following recent work on Parkinson's disease detection from speech in English [20], we adopted the following prosodic and acoustic features: MFCC, Pitch, Jitter (absolute, relative, rap, and PPQ5), Shimmer (absolute, absolute dB, relative, APQ3, and APQ5), Harmonicity. We concatenated these prosodic and acoustic features into a single, di-dimensional, feature map.

We applied chunking to the feature maps resulting from features extraction using non-overlapping windows of 4 s. We applied padding to make sure that all windows ended up composed of 4 s data. We extended both sides of the feature map (before the chunking step), replicating the value on the border. We repeated the values so that the input sequence of features could be decomposed into an integer number of chunks. For all the considered input features, we applied standard scaling, computing mean and variance of the individual features vectors composing all the feature maps.

After chunking, we obtained 538 samples, 315 from people in healthy conditions and 223. We applied minority oversampling to balance the data set.

Concerning the CNN, we used a standard stack of 1D convolutional layers with non-linear activation as hidden transformation, followed by the vectorization operation and a final linear classification layer. As anticipated in Section 3.3, we considered flattening, GAP and GMP as vecotrisation transformations. For each input feature-vectorisation algorithm pair, we searched for the best hyperparameters configuration of the CNN using 5-fold cross-validation. We considered 1, 2, or 3 convolutional blocks and either 512 or 1024 output channels for the convolutions. We used a constant kernel width of 3 for all configurations. We trained the CNN using the Adam optimiser, in the cross-validation step we considered as learning rates 10^{-3} and $5 \cdot 10^{-4}$. To prevent overfitting we used dropout with a probability of 10%

5.2. Evaluation approach

To ensure a correct evaluation of the model performances, we split the audio segments into train and test, with a 75%-25% split. We used the same training and testing subsets with each of the proposed models (i.e., pipeline implementations). For each model, we computed the common metrics used in machine learning for classification and information retrieval problems, defined starting from the confusion matrix.

Table 1 - Confusion matrix.

	Predicted positive	Predicted negative
Labelled positive	TP	FN
Labelled negative	FP	TN

Referring to the confusion matrix in Table 1, we introduce the following definitions:

- $TP =$ True Positive (i.e., positive values correctly predicted as such)
- $TN =$ True Negative (i.e., negative values correctly predicted as such)
- $FP =$ False Positive (i.e., negative values predicted as positive)
  $FN =$ False Negative (i.e., positive values predicted as negative)

Given that this is a Parkinson's disease detection problem, we associate the positive class with the disease condition and the negative class with the healthy condition.

To assess the quality of the trained classification models, we computed the following metrics:

- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- precision $= \frac{TP}{TP+FP}$ (i.e., positive predictive value)
- recall $= \frac{TP}{TP+FN}$ (i.e., sensitivity, hit rate, true positive rate)
- specificity $= \frac{TN}{TN+FP}$ (i.e., selectivity, negative class recall true negative rate)
- F_1-score $= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ (i.e., sensitivity, hit rate, true positive rate)
- AUC (Area Under the Curve of the Receiver Operating Characteristic)

5.3. Results and comments

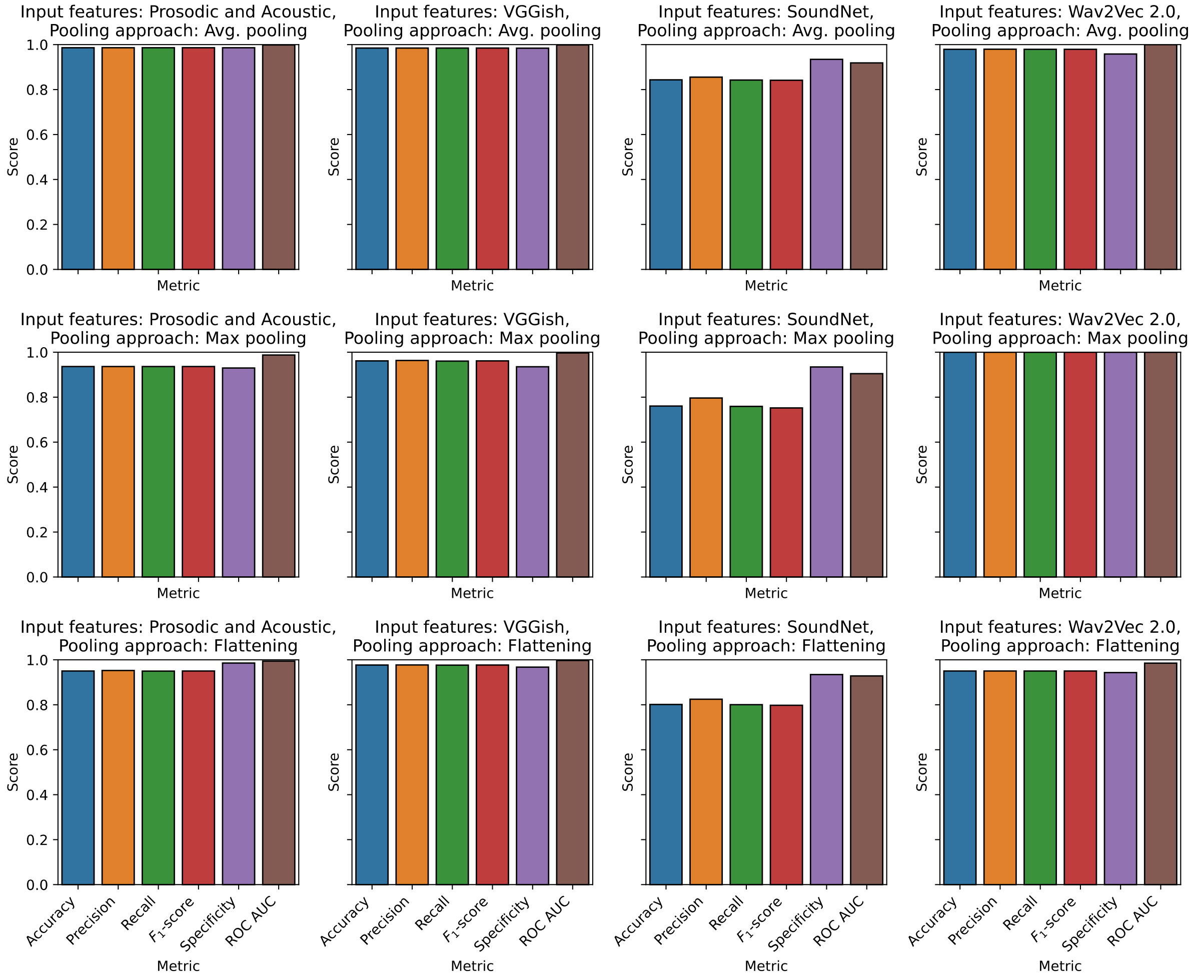

Figure 5.1. Results of the different tested configurations. Each row corresponds to a different pooling (vectorization) approach, each column corresponds to an input feature.

We reported the measured metrics on the test split in Figure 5.1. All models managed to achieve good performances despite the reduced data set size: in most cases we achieved scores for all metrics > 95%. All features showed to be independent from the pooling approach, reaching similar results across the different vecotrisation algorithms.

Concerning deep learning features, VGGish and Wav2Vec 2.0 achieve the overall best results. On the other side, SoundNet produced the worst results. In all cases where the classifiers produce worse results, we can notice that the recall score is lower than the precision one. Thus, we can hypothesize that, in those cases, the unbalance in the number of negative samples (corresponding to healthy patients) influenced negative the model, causing the increase of false negatives.

Interestingly, the handcrafted features perform comparably to VGGish and Wav2Vec 2.0. In fact, despite the small data set, we managed to achieve almost a perfect score (which is not possible). This hints that the few handcrafted features encode very useful information for the task.

6. Conclusion

In this chapter, we approached a speech analysis problem, Parkinson's disease detection from voice, using a machine learning pipeline. We evaluated different feature extraction approaches, comparing traditional prosodic and acoustic features against features computed by deep neural networks. We used the extracted features to fit an CNN classifier. All the experiments were conducted on a relatively small data set of audio clips collected from Telugu speakers. We achieved equally good results using both deep features and, surprisingly, with the handcrafted features. The reported scores are in line with those achieved on bigger data sets, showing that even with low resources deep learning models can yield good generalization capabilities. Nevertheless, handcrafted features showed to be capable of yielding valid results, comparable to those of the deep models, despite the reduced data set size. This hints that deep learning solutions for audio processing still need to become an irreplaceable tool. For the sake of reproducibility, we are sharing the source code via *GitHub*[1].

Concerning future direction, we are willing to explore different learning paradigms to improve the detection results. On one side we are considering exploiting regularities within data and see if unsupervised learning may lead to better results. Ideally, the similarities and dissimilarities between samples may be used to feed a clustering algorithm, allowing, possibly, to group automatically samples from healthy patients and patients affected by Parkinson's disease. On the other side, we are considering anomaly detection approaches. In fact, given that there are many more available samples of speech from people in healthy conditions, it would be possible to detect samples of speech from Parkinson's disease patients as outliers to the distribution of the regular data.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li e P. J. Liu, «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,» *J. Mach. Learn. Res.*, vol. 21, p. 140:1–140:67, 2020.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever e D. Amodei, «Language Models are Few-Shot Learners,» in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[3] A. Baevski, Y. Zhou, A. Mohamed e M. Auli, «wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,» in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey e I. Sutskever, «Robust Speech Recognition via Large-Scale Weak Supervision,» *OpenAI blog*, p. 28, 2022.

[1] https://github.com/vincenzo-scotti/voice_analysis_parkinson

[5] L. Wan, Q. Wang, A. Papir e I. Lopez-Moreno, «Generalized End-to-End Loss for Speaker Verification,» in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018.

[6] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark e R. A. Saurous, «Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron,» in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018.

[7] A. Favaro, L. Sbattella, R. Tedesco e V. Scotti, «ITAcotron 2: Transfering English Speech Synthesis Architectures and Speech Features to Italian,» in *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, Trento, 2021.

[8] A. Favaro, L. Sbattella, R. Tedesco e V. Scotti, «ITAcotron 2: the Power of Transfer Learning in Expressive TTS Synthesis,» in *Analysis and Application of Natural Language and Speech Processing*, M. Abbas, A cura di, Cham, Springer International Publishing, 2022, p. 1–20.

[9] V. Scotti, F. Galati, L. Sbattella e R. Tedesco, «Combining Deep and Unsupervised Features for Multilingual Speech Emotion Recognition,» in *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part II*, 2020.

[10] J. Yosinski, J. Clune, Y. Bengio e H. Lipson, «How transferable are features in deep neural networks?,» in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014.

[11] T. Giannakopoulos, «pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis,» *PLOS ONE*, vol. 10, pp. 1-17, December 2015.

[12] V. Chernykh, G. Sterling e P. Prihodko, «Emotion Recognition From Speech With Recurrent Neural Networks,» *CoRR*, vol. abs/1701.08071, 2017.

[13] Y. Aytar, C. Vondrick e A. Torralba, «SoundNet: Learning Sound Representations from Unlabeled Video,» in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

[14] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss e K. W. Wilson, «CNN architectures for large-scale audio classification,» in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017.

[15] S. Schneider, A. Baevski, R. Collobert e M. Auli, «wav2vec: Unsupervised Pre-Training for Speech Recognition,» in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, 2019.

[16] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. A. Ciccarelli e D. D. Mehta, «Segment-dependent dynamics in predicting parkinson's disease,» in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015.

[17] B. Karan, S. S. Sahu e K. Mahto, «Parkinson disease prediction using intrinsic mode function based features from speech signal,» *Biocybernetics and Biomedical Engineering,* vol. 40, p. 249–264, 2020.

[18] W. Rahman, S. Lee, M. S. Islam, V. N. Antony, H. Ratnu, M. R. Ali, A. Al Mamun, E. Wagner, S. Jensen-Roberts, E. Waddell e others, «Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study,» *Journal of medical Internet research,* vol. 23, p. e26305, 2021.

[19] S. Kurada e A. Kurada, «Poster: 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, December 16-18, 2020,* 2020.

[20] A. A. Toye e S. Kompalli, «Comparative Study of Speech Analysis Methods to Predict Parkinson's Disease,» *CoRR,* vol. abs/2111.10207, 2021.

[21] A. Favaro, S. Motley, Q. M. Samus , A. Butala, N. Dehak, E. S. Oh and L. Moro-Velazquez, "Artificial Intelligence Tools to Evaluate Language and Speech Patterns in Alzheimer's Disease," *Alzheimer's & Dementia,* vol. 18, no. S2, 2022.

[22] A. Favaro, C. Montley, M. Iglesias, A. Butala, E. S. Oh, R. D. Stevens, J. Villalba, N. Dehak e L. Moro-Velasquez, «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, 2022.

[23] H. Jaeger, D. Trivedi e M. Stadtschnitzer, *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls,* Zenodo, 2019.

[24] G. Dimauro e F. Girardi, *Italian Parkinson's Voice and Speech,* IEEE Dataport, 2019.

[25] G. Dimauro, D. Caivano, V. Bevilacqua, F. Girardi e V. Napoletano, «VoxTester, software for digital evaluation of speech changes in Parkinson disease,» in *2016 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016, Benevento, Italy, May 15-18, 2016*, 2016.

[26] I. J. Goodfellow, Y. Bengio e A. C. Courville, Deep Learning, MIT Press, 2016.

[27] M. J. Zaki e W. Meira, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2 a cura di, Cambridge University Press, 2020.

[28] F. He, S.-H. C. Chu, O. Kjartansson, C. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. Johny, M. Jansche, S. Sarin e K. Pipatsrisawat, «Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems,» in *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 2020.

[29] J.-M. Valin, «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, August 29-31, 2018*, 2018.

[30] M. Lin, Q. Chen e S. Yan, «Network In Network,» in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[31] G. Dimauro, V. D. Nicola, V. Bevilacqua, D. Caivano e F. Girardi, «Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System,» *IEEE Access,* vol. 5, p. 22199–22208, 2017.