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

Claudio Ferrante and **Vincenzo Scotti**
DEIB, Politecnico di Milano – Via Golgi 42, 20133, Milano (MI), Italy
claudio1.ferrante@mail.polimi.it vincenzo.scotti@polimi.it

Traditionally, *speech analysis* has always relied on a set of very informative features like *(Mel) spectrogram* or *Mel Frequency Cepstral Coefficients (MFCC)* to solve many different problems and build incredible speech powered applications (Jurafsky and Martin, 2009). In this sense, speech analysis distinguished itself from other areas, like *computer vision*, where deep learning has become a pervasive framework. However, recently, deep learning models for the extraction of acoustic features have allowed improving significantly the state of the art in many speech-related applications, like *Automatic Speech Recognition (ASR)* (Baevski et al., 2020; Radford et al., 2022), *speaker identification* (Wan et al., 2018), or *speech emotion recognition* (Scotti et al., 2020). Basically, these models are *deep neural networks* trained in an *unsupervised* or *self-supervised* (Aytar et al., 2016; Hershey et al., 2017; Baevski et al., 2020) way on large collections of acoustic data (not necessarily speech). Starting from the features computed by these deep neural network models, we can build classification and regression models to solve speech analysis-related problems with improved results.

With our work, we focus the analysis on the cross-lingual transferability of these deep learning features for speech analysis. The idea is to understand whether and how well a classification model trained on deep acoustic features in a source language works in a new target language. The problem we are dealing with is thus a *domain adaptation* problem (Redko et al., 2019). We selected Parkinson's disease recognition as the use case of our experiments: the goal is to train a binary classifier that discriminates from speech between healthy patients and patients affected by Parkinson's disease.

We used two data sets for our experiments: one in English and one in Telugu, the latter served as an under-resourced language example. We experimented using both languages first as the source domain data and as the target domain data in two sets of experiments. The English data set is the *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls* (Jaeger et al., 2019). It contains both spontaneous and read audio clips from both Parkinson's disease patients and healthy patients, respectively it contains 266 clips of healthy persons and 199 clips of persons affected by Parkison's disease. The Telugu data set, instead, is a private collection of audio clips from Parkinson's disease patients (81 clips); we combined it with samples taken from the Telugu split of *Open SLR* (He et al., 2020), a data set for speech recognition (we subsampled 200 clips from it). The latter language motivated our work; in fact, the scarcity of data for some problems or languages usually forces us to adapt solutions from different domains.

Building a classifier to recognise Parkinson's disease from the speech is not a new problem. This task has already been addressed using machine learning-based techniques (Williamson et al., 2015; Karan et al., 2020; Rahman et al., 2021; Kurada and Kurada, 2020; Toye and Kompalli, 2021). Usually, these approaches involve the extraction of acoustic and prosodic features (mainly MFCC, *Jitter*, *Shimmer*, and *Pitch*), which allows the building of very discriminative models. In some cases, these features are further processed through dimensionality reduction transformations to keep only the most relevant components of the vectors encoding the audio clips to classify.

Very recent results already explored the role of deep learning features to learn a classifier for this Parkinson's disease detection problem (Kurada and Kurada, 2020), reaching more than 80% recognition accuracy on the same English data set we considered and outperforming the other considered classifiers based on acoustic features. Other works started focusing on building classifiers for other languages (Toye and Kompalli, 2021): their results showed that acoustic and prosodic features could be used to build high-performing classifiers (reaching more than 90% recognition accuracy) on English and Italian. However, none of these previous works analysed the effects of directly transferring trained classification models for Parkison's disease recognition across languages (usually models are trained from scratch in the new language, which isn't always possible due to data availability problems) or the effects of applying domain adaptation on these classifiers to see how it affects the performances in the target language, without explicit retraining. With this work, we are interested in experimenting with classification models to tackle these two analyses.

The classification pipeline we propose is completely based on data-driven techniques. The first step we apply is denoising through *RNNoise* (Valin, 2018), a deep neural network for voice enhancement and noise suppression. We process the denoised audio directly with one of the selected deep features models, which yields a sequence of feature vectors. Each sequence vector covers a specific time window of the input signal. Since the final classification step of machine learning models expect information to be encoded into a single vector, we tested different approaches to convert the sequence of feature vectors into a single vector: *flattening* (or unrolling), *Global Average Pooling (GAP)* (Lin et al., 2014), and *Global Max Pooling (GMP)* (Lin et al., 2014). Finally, we used a *Feed Forward Neural Network* to perform the classification task from the extracted feature vectors.

As premised, we considered also a technique for *unsupervised domain adaptation*, *Deep CORAL* (Sun and Saenko,

2016), as an alternative path of our pipeline. We re-trained the classification network adding the Deep CORAL objective on the covariance of the final projections before classification for adaptation. We considered this adaptation step to understand how robust the classifier is with and without explicit adaptation.

In our work, we considered three different models for computing deep acoustic features: *Wav2Vec* (Baevski et al., 2020), *VGGish* (Hershey et al., 2017), and *SoundNet* (Aytar et al., 2016). The former was specifically designed for speech analysis problems and is used as input for state-of-the-art ASR models. The other two instead are actually mode generic models though for acoustic analysis not necessarily aimed at speech. VGGish was trained on a large audio classification data set containing many labels. SoundNet, on the other hand, was trained to predict from the audio track of video clips the pseudo labels generated from object recognition and scene recognition using a neural network that processed the images of the video clips. Additionally, we trained some classification models using traditional acoustic and prosodic features. We selected the set of features from recent works on the same problem and data set (Toye and Kompalli, 2021).

The results on individual languages we obtained are in line with previous works. Even with few samples, with Telugu, we got $> 90\%$ accuracy with some feature and pooling configurations. In English, we reached $> 90\%$ accuracy as well. We achieved these results using either traditional or deep features, showing that for this kind of speech classification problems, transferring deep features is not necessary to achieve the best results. Concerning cross-lingual results, we noticed that in some cases, without explicit adaptation, many classifiers still achieve reasonable results ($\geq 70\%$ accuracy) when applied to the other (target) language. From the results of domain adaptation, we noticed that when the zero-shot behaviour of the unadapted classifier produced low scores on the target language, adaptation helped improve the performances, lowering the scores on the source language though. The results of adaptation are mixed, and a more extensive analysis of these techniques is necessary to understand how to achieve the best from these approaches. As for now re-training and zero-shot model transfer seem to be the most effective solutions. For completeness and reproducibility, we share the reference to the repository with the source code[1].

References

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 892–900, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6494–6503. European Language Resources Association, 2020.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin 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*, pages 131–135. IEEE, 2017.

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

Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009.

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

Sruthi Kurada and Abhinav 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*, pages 9–11. IEEE, 2020.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *OpenAI blog*, page 28, 2022.

Wasifur Rahman, Sangwu Lee, Md Saiful Islam, Victor Nikhil Antony, Harshil Ratnu, Mohammad Rafayet Ali, Abdullah Al Mamun, Ellen Wagner, Stella Jensen-Roberts, Emma Waddell, et al. Detecting parkinson disease using a web-based speech task: Observational study. *Journal of medical Internet research*, 23(10):e26305, 2021.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. *Advances in domain adaptation theory*. Elsevier, 2019.

Vincenzo Scotti, Federico Galati, Licia Sbattella, and Roberto Tedesco. Combining deep and unsupervised features for multilingual speech emotion recognition. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part II*, volume 12662 of *Lecture Notes in Computer Science*, pages 114–128. Springer, 2020.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, pages 443–450, 2016.

Adedolapo Aishat Toye and Suryaprakash Kompalli. Comparative study of speech analysis methods to predict parkinson's disease. *CoRR*, abs/2111.10207, 2021.

Jean-Marc 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*, pages 1–5. IEEE, 2018.

Li Wan, Quan Wang, Alan Papir, and Ignacio 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*, pages 4879–4883. IEEE, 2018.

James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Joseph Perricone, Satrajit S. Ghosh, Gregory A. Ciccarelli, and Daryush 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*, pages 518–522. ISCA, 2015.

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