ECSE-552 Project Proposal

Group 1: Rubert Martin & Achaebe Parker & Graham Smith & Max Henry & Emmanuel Wilson

I. Background

Our chosen task is language identification based on speech signals [1, 2, 3, 4]. We propose using convolutional neural nets (CNNs) which are a favourite for audio learning [5]. (Specifically, we will use an evolution of CNNs that is a residual network). Given recent near-perfect performance on this task [4], we also propose a secondary challenge: to leverage self-supervised pretraining and transfer learning to train a performant model on relatively *few data*. Pre-training strategies include prototype networks [6] and auto-encoders [7].

II. Goals

- Create a classifier that can identify at least one spoken language from an audio dataset.
- Determine the limits of data amounts and their effect on model training.
- Get familiarized with ResNet architecture.
- Understand how time-frequency representations of audio data affect model training.

III. Methodology

We will feed spectrograms to a ResNet50 network pretrained on images, as per [3]. We will utilize the Librosa library[1] to generate the spectrograms, and build the network using PyTorch. After establishing a shallowly-trained baseline, we will explore the proposed pretraining techniques. In order to ensure all individuals have access to the same computational platform and resources, our models will be trained and validated using Google Colab.

Classification performance will be evaluated using multiclass accuracy, following precedent [3, 4, 6]. Other measures like AUROC may also be considered.

IV. Dataset

Sound clips for training and validation will be gathered from VoxForge[2], an open-source initiative with user-supplied examples. The most popular languages by submission are: English (6319), French (2260), Spanish (2248), German (1419) and Italian (1060). Each submission contains a variety of prompts which are spoken and recorded by the user. All recordings are saved as .wav files in a variety of sampling rates and bit-depths, ranging from 8–48kHz and 16–32 bits respectively. Included in each submission is a transcription of the prompts supplied to the users, as well as information on the recording setup.

[1] https://librosa.org/

[2] http://www.voxforge.org/

V. Potential pitfalls

Voxforge provides an open-source user-supplied dataset, so there is no standardization in recording methodologies. This means artifacts could be introduced into the audio from the data capturing process which might influence the classifier. Potential mic clipping may add frequency content to the signal in a way similar to a phoneme from another language, causing the sample to be misclassified. This issue could be mitigated by pre-processing the data to eliminate the artifact, or sorting the data to identify problematic training examples and creating a separate training stage for them after the more ideal data has been processed.

VI. Roles

Given the professional background and expertise of each team member, Max and Graham will contribute with digital audio processing techniques, Emmanuel will assist on machine learning architectures and Rubert and Achaebe will help with the mathematical formulation of the problem. All members will develop the model and provide input regarding experimental design and data analysis.

References

[1] Gregoire Montavon. "Deep learning for spoken language identification". In: *NIPS Workshop on deep learning for speech recognition and related applications*. 2009.

[2] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. "Automatic language identification using deep neural networks". In: *ICASSP*. 2014, pp. 5337–5341.

[3] Shauna Revay and Matthew Teschke. "Multiclass language identification using deep learning on spectral images of audio signals". In: *arXiv:1905.04348* (2019).

[4] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. "Towards learning a universal non-semantic representation of speech". In: *arXiv:2002.12764* (2020).

[5] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. "CNN architectures for large-scale audio classification". In: *ICASSP*. 2017, pp. 131–135.

[6] Jordi Pons, Joan Serrà, and Xavier Serra. "Training neural audio classifiers with few data". In: *ICASSP*. 2019.

[7] Dor Bank, Noam Koenigstein, and Raja Giryes. "Autoencoders". In: *arXiv:2003.05991* (2020).