Speaker identification with deep neural networks

Alfredo Méndez
amendezp4@gmail.com

Juan Pablo Rodríguez
juanpablordz94@gmail.com

Motivation

There is a great research effort in looking for medical solutions for **Alzheimer's disease**, while significantly less in creating solutions for caregiving post-diagnosis. The motivation of this project is to provide patients with a tool to recognize familiar people in common situations.

The solution implements a t**ext independent speaker recognition** system using deep learning. As input, the model receives spoken utterances to output a predicted identity profile.

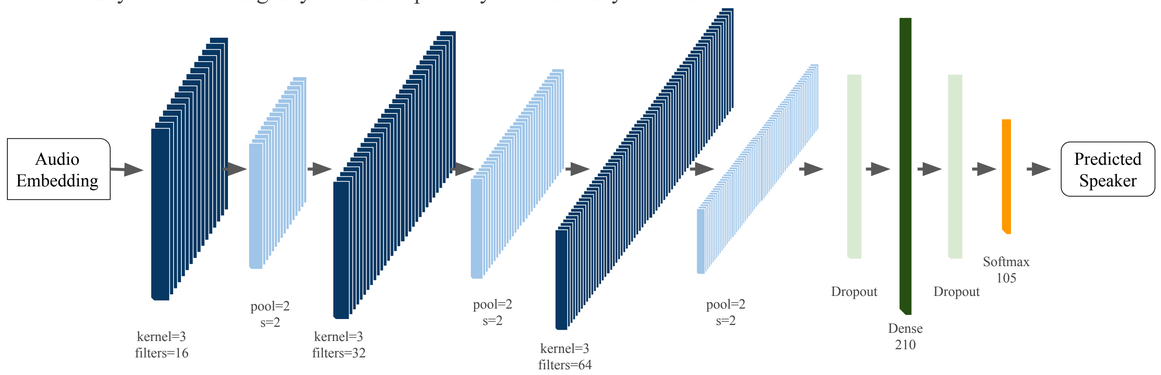
Models

DNN Model

2 FC Layers - Sigmoid activation functions

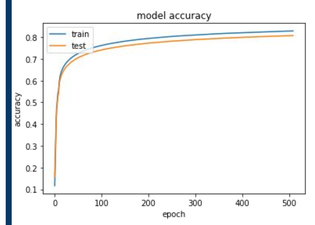
CNN Model

3 Conv. Layers - 3 Pooling Layers - 2 Dropout layers - 1FC Layer - ReLu activation functions

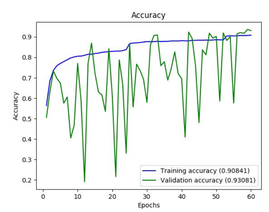


Results

Model	Train Accuracy	Test Accuracy	Train Loss	Test Loss
DNN	78.05%	75.88%	0.6951	0.7421
CNN	90.84%	93.08%	0.2956	0.2420



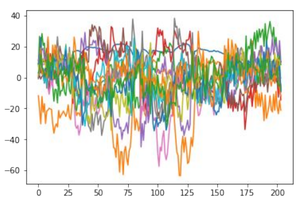
DNN Model Accuracy



CNN Model Accuracy

Data

VCTK Corpus composed by speech data uttered by **109 english speakers.** Every speaker reads around **400 sentences** that were specifically selected to maximize contextual and phonetic coverage. Each speaker reads a different set of sentences, which is good for a text-independent model.



Discussion

- **DNN Model** has low variance and high bias. A different architecture could help improve.
- **CNN Model** presented high variance and low bias on the training set. This suggests that there is still overfitting. The model did not converged on 80 epochs, adjusting learning rate and regularization techniques could lead to better performance.
- There is **no sufficient data to choose a "best" model.**

Future Work

1. Implement a **hyperparameter tuning** module.
2. Tain CNN until convergence.
3. Perform **data augmentation** to add background noise.
4. Explore the performance in RNN and Residual architectures.
5. **Test with Alzheimer's patients** in common settings. Develop a proof of concept product.

References

1. Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. (2017). Deep speaker: an end-to-end neural speaker embedding system. arXiv preprint arXiv:1705.02304.
2. Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. In 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP), pages 1–6. IEEE.
3. Torfi, A., Dawson, J., and Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In 2018 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE.
4. Centre for Speech Technology Research. VCTK Corpus.
5. Philippe Remy - Deep Speaker
6. Manish Pandit - Speaker Recognition