

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

1 Introduction

Given audio recordings of ten people reading the digits 0 to 9, we wish to train a computer to take the `.wav` files and accurately classify the digit being spoken. We follow [1] by converting the audio files into mel-frequency spectral coefficients, producing an image that can be inputted into a 2D convolutional neural network (CNN). We also compare this CNN with a recurrent neural network (RNN).

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1533–1545, 2014.