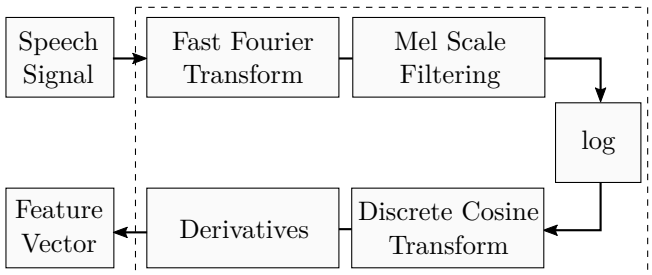
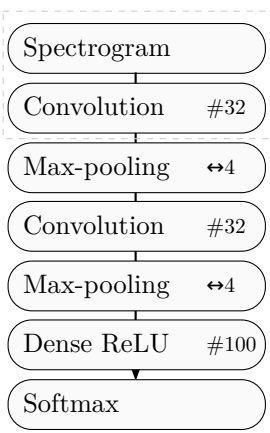
Convolutional Neural Networks and Algebraic Scale Invariance for Speech Classification

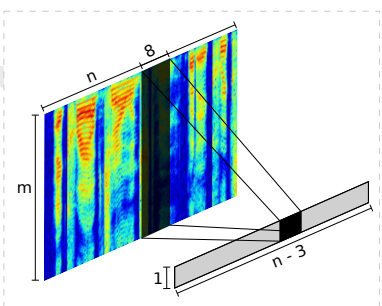
MFCC



Current models use complex human-enginered MFCC features for modelling.

The Dieleman Network





The Dieleman network can hopefully learn the MFCC features or something better.

Scale Invariant Regularization

$$\mathcal{R}(s) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial P(C_{i,k}|s(x_i,\alpha),w)}{\partial \alpha}\bigg|^2_{\alpha=0}$$

Scale invariant

$$s(x,\alpha) = (1+\alpha)x \qquad \mathcal{R} = \frac{1}{N}\sum_{i=1}^{N} (\nabla_x P(C_{i,k}|x_i,w)\cdot x_i)^2$$

Offset invariant

$$s(x,\alpha) = x+\alpha \qquad \mathcal{R} = \frac{1}{N}\sum_{i=1}^{N} (\nabla_x P(C_{i,k}|x_i,w)\cdot \mathbf{1})^2$$

Dieleman Results

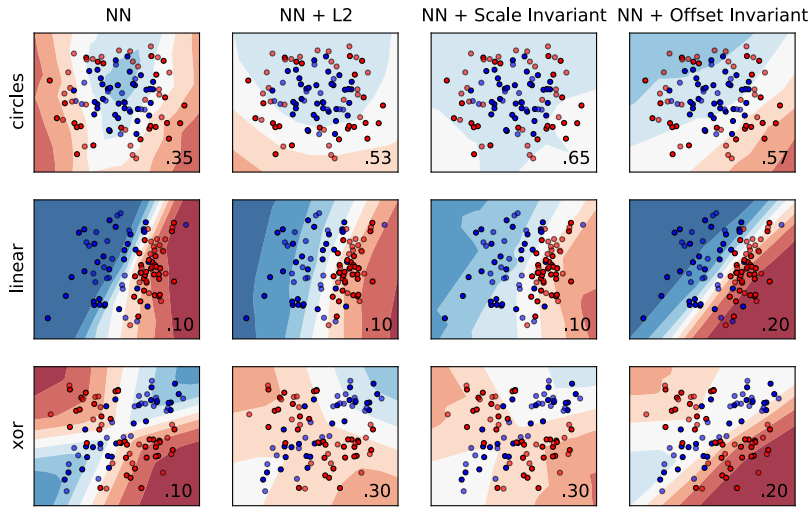
	TIMIT	ELSDSR
Baseline	0.354	0.465
Logistic on mean	0.094 ± 0.012	0.030 ± 0.007
GMM on MFCC	0.192 ± 0.024	0.140 ± 0.019
Dieleman	0.093 ± 0.012	0.026 ± 0.006
Dieleman +L2	0.114 ± 0.013	0.036 ± 0.016
Dieleman +Scale	0.111 ± 0.015	0.022 ± 0.006
Dieleman +Offset	0.107 ± 0.008	0.027 ± 0.014

Missclassification rate on sex classification.

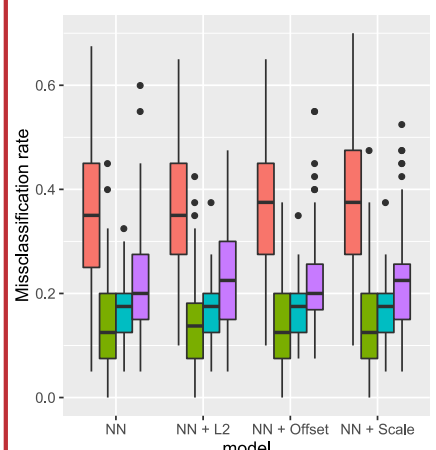
	TIMIT	ELSDSR
Baseline	0.988	0.957
Logistic on mean	0.796 ± 0.046	0.338 ± 0.043
GMM on MFCC	0.836 ± 0.020	0.391 ± 0.023
Dieleman	0.965 ± 0.021	0.570 ± 0.029
Dieleman +L2	0.944 ± 0.020	0.552 ± 0.045
Dieleman +Scale	0.973 ± 0.007	0.640 ± 0.110
Dieleman +Offset	0.971 ± 0.006	0.628 ± 0.117

Missclassification rate on speaker classification.

Regularization Analysis



Contours of probability function on 3 synthetic datasets using extreme regularization parameters.



Missclassification rate boxplot using optimized regularization parameters.

DTU Compute
Institut for Matematik og Computer Science