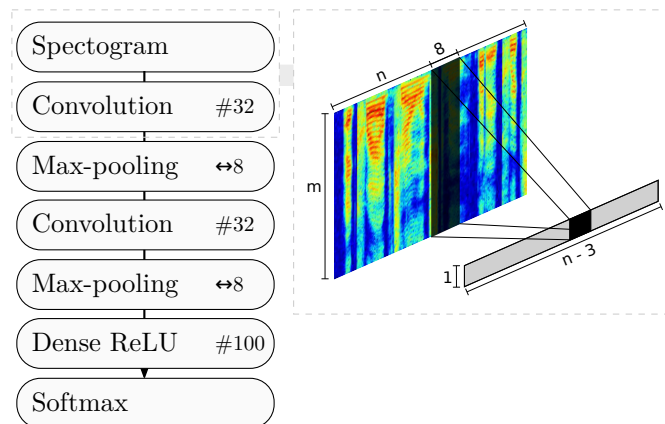# Convolutional Neural Networks and Algebraic Scale Invariance for Speech Classification

The Dieleman Network

- Spectogram
- Convolution   #32
- Max-pooling   ↔8
- Convolution   #32
- Max-pooling   ↔8
- Dense ReLU   #100
- Softmax

Scale Invariant Regularization

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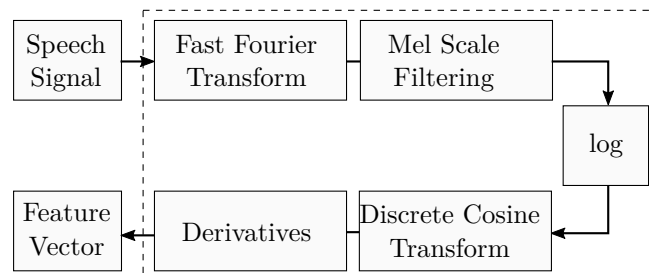
Scale invariant

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

Offset invariant

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

MFCC

Speech Signal → Fast Fourier Transform → Mel Scale Filtering → log → Discrete Cosine Transform → Derivatives → Feature Vector

Mel-Frequency Cepstral Coefficients computation steps

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

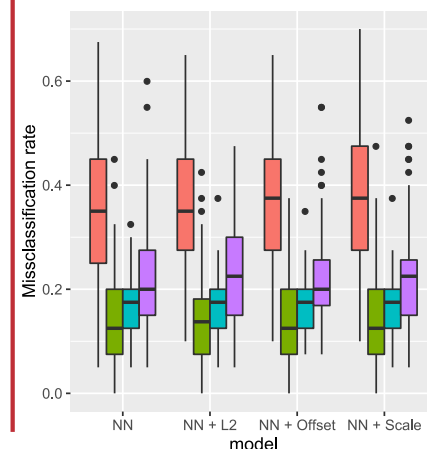*misclassification rate on sex classification.*

	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

misclassification rate on speaker classification.

Regularization Analysis
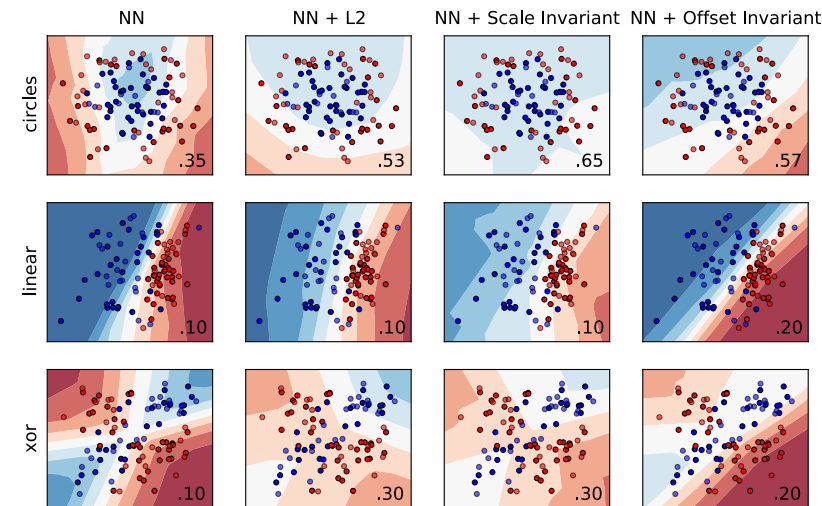
contours of probability function on 3 synthetic datasets using extreme regularization parameters.

misclassification rate boxplot using optimized regularization parameters.