Speaker Recognition Using Deep Neural Networks

Harald Stiff **Carl Jernbäcker**

Royal Institute of Technology

Problem Formulation

- Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ where each $\boldsymbol{x}_i \in \mathbb{R}^n$ is an auditory feature vector belonging to a corresponding onehot encoded speaker label $\boldsymbol{y}_i \in \mathbb{R}^m$.
- Can we define a neural network estimator $\hat{\boldsymbol{y}} = f(\boldsymbol{x}, \Theta)$ such that $\forall(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}, \;\; f(\boldsymbol{x}_i, \Theta) \approx \boldsymbol{y}_i$?

Method

Theory

- We consider the N-layered neural network estimator in Figure 1

$$f(\boldsymbol{x}, \Theta) = \text{SoftMax}(\boldsymbol{W}^N \sigma(\boldsymbol{W}^{N-1}\sigma(\ldots \sigma(\boldsymbol{W}^1\boldsymbol{x} + \boldsymbol{b}^1)) + \boldsymbol{b}^{N-1}) + \boldsymbol{b}^N)$$

where $\Theta = \{\boldsymbol{W}^i, \boldsymbol{b}^i\}_{i=1}^{N}$ contains the parameters of the network.

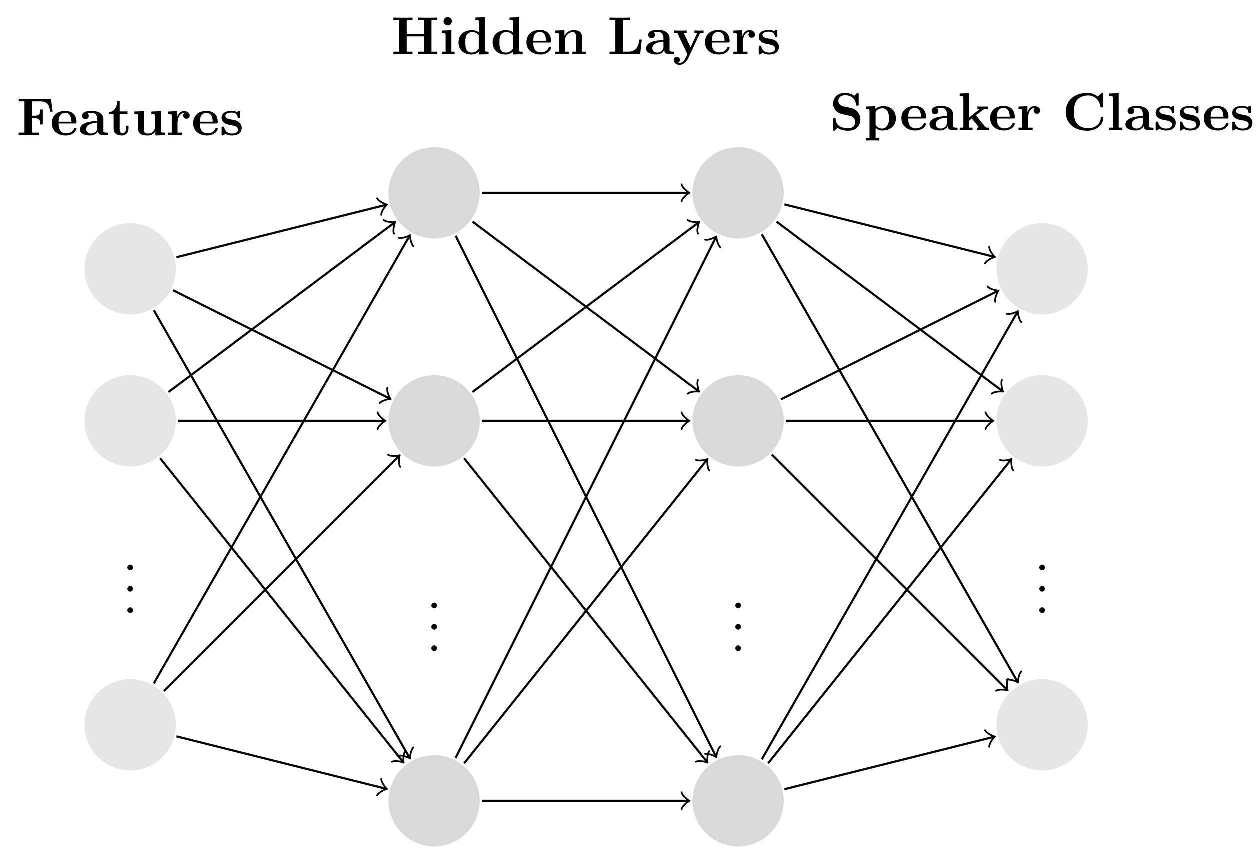

Figure 1: **An artficial neural network.**

- The parameters in Θ are updated using an ADAM optimizer provided the gradients

$$\left\{\frac{\partial J(\mathcal{B}, \Theta)}{\partial \boldsymbol{W}^i}, \frac{\partial J(\mathcal{B}, \Theta)}{\partial \boldsymbol{b}^i}\right\}_{i=1}^{N}.$$

where

$$J(\mathcal{B}, \Theta) = \frac{1}{n}\sum_{i=1}^{n} -\log(\boldsymbol{y}_i^T \hat{\boldsymbol{y}}_i)$$

and $\mathcal{B} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ is a stochastic batch of features and labels.

- We define the m voted neural network estimator as

$$\hat{\boldsymbol{y}}^{(m)} = \frac{1}{m}\sum_{k=1}^{m} f(\boldsymbol{x}_k, \Theta)$$

where m feature vectors belonging to the same class are used for one estimation.

Feature Collection

- We extract normalized MFCC and mel filter bank features from 112 speakers in the TIDIGITs dataset in according to Figure 2.

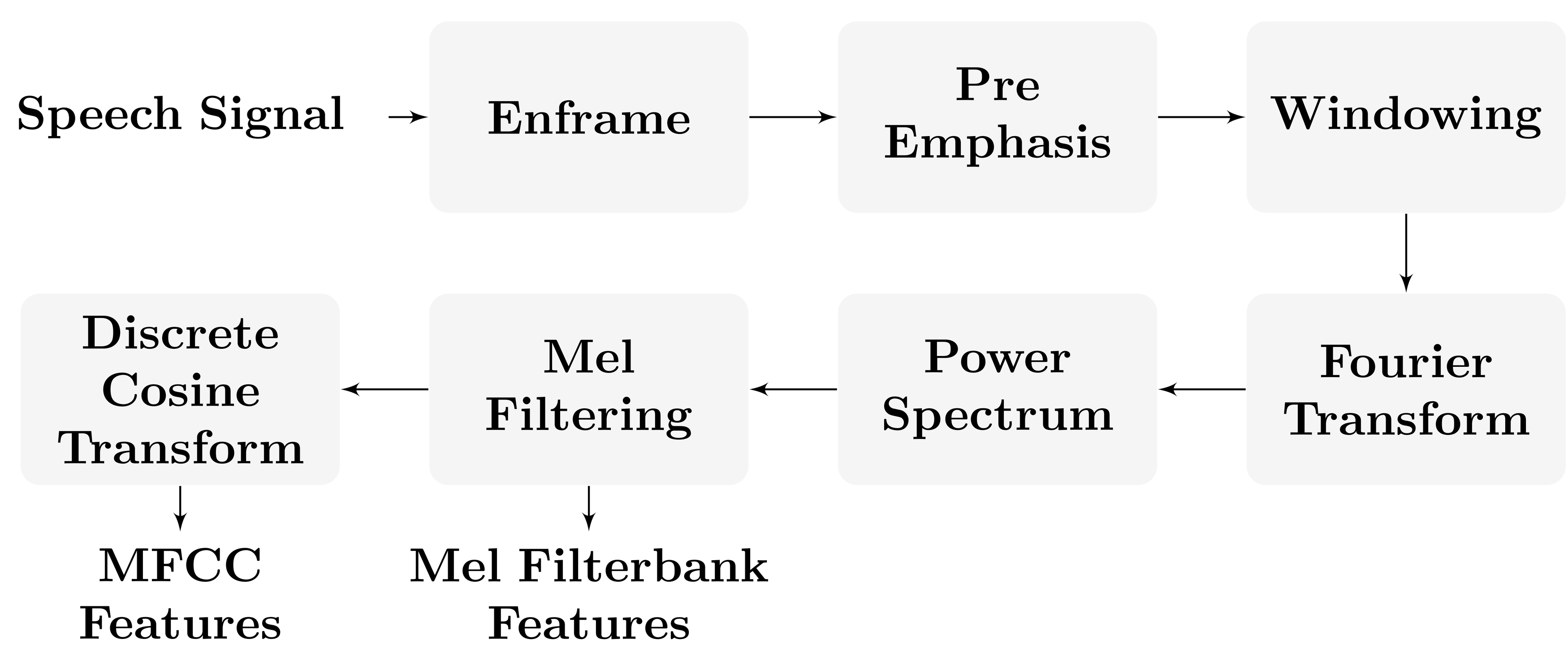

Figure 2: **Feature collection flow chart**

Results

- The best results are shown in Tables 1-2 where M, F, MF correspond to networks trained on only male speakers, female speakers and both male and female speakers respectively.
- Each hidden layer in the networks has 256 nodes
- Figures 3-4 are plots comparing the performance of the estimator using a varied amount of votes.
- Figure 5 show accuracy and loss plots with respect to the number of epochs used in training.

Table 1: Mel Filter Bank Features

Model	h-layers	gender	acc	loss
1	3	M	43.4	2.18
2	3	F	46	2.06
3	3	M&F	39.3	2.5

Table 2: MFCC Features

Model	h-layers	gender	acc	loss
1	5	M	34	2.55
2	5	F	31	2.71
3	5	M&F	26,4	3.1

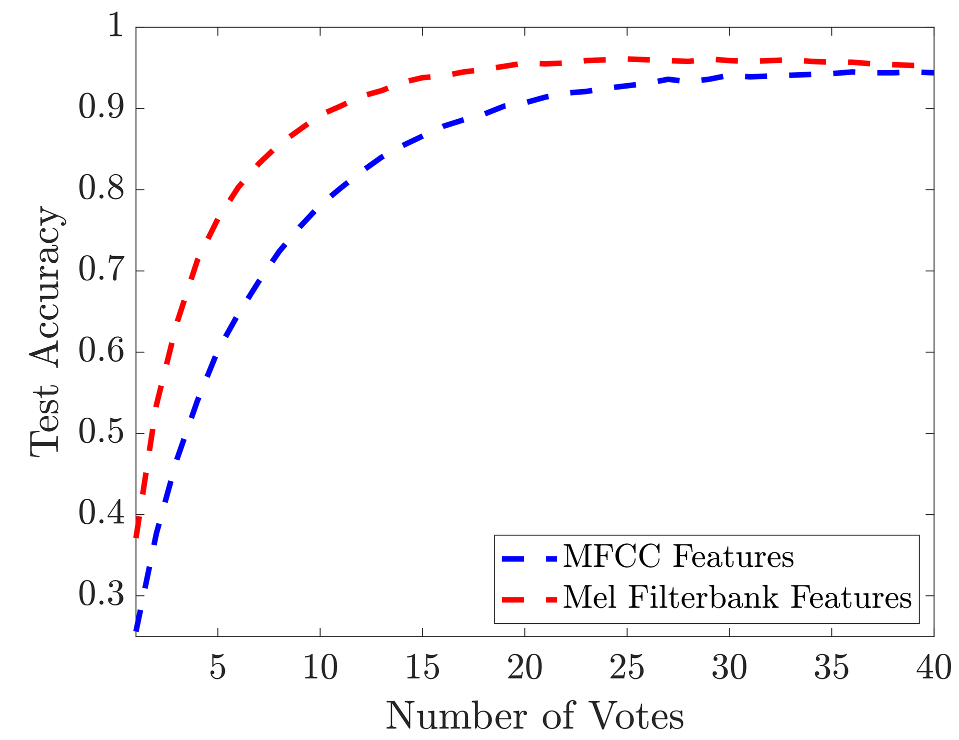

Figure 3: **Accuracy of different estimators $\hat{y}^{(m)}$ using model 3 in Table 1.**

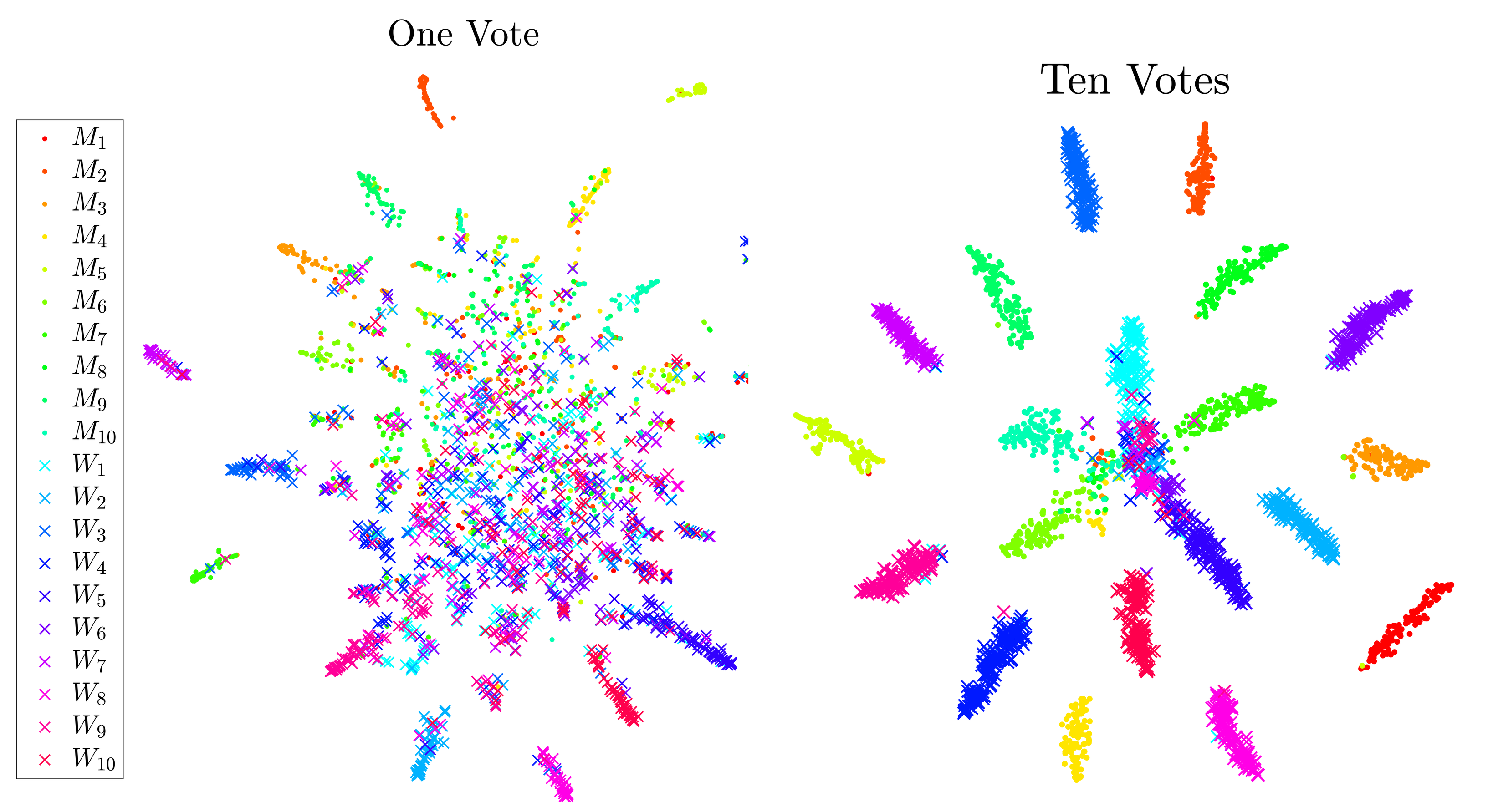

Figure 4: **PCA cluster plots of classifications $\hat{y}^{(1)}$ and $\hat{y}^{(10)}$ using model 3 in Tables 1-2. Only classifications of features belonging to 10 male and 10 female speakers are shown.**

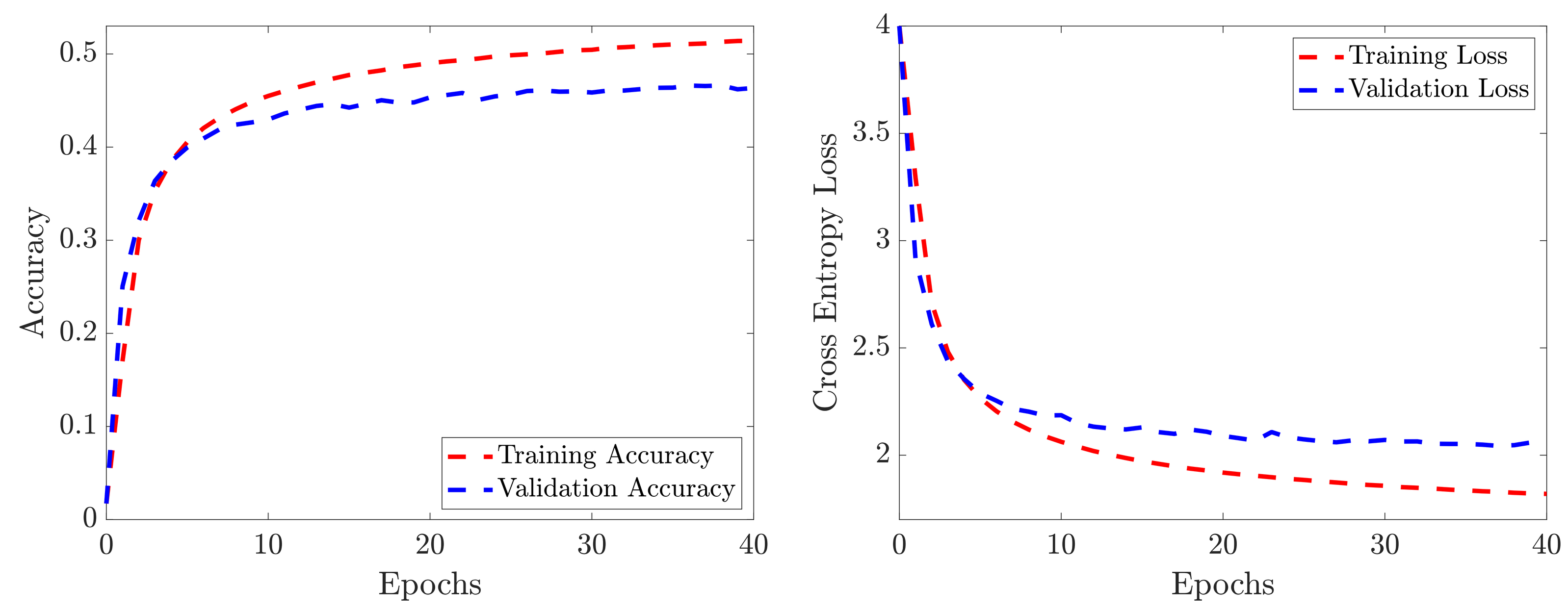

Figure 5: **Accuracy and loss plots for model 2 in Table 1.**