HEIG-VD

Report

MLG Labo2: Voice Recognition Experiments

Authors :
Basile Châtillon
Valentin Finini

April 25, 2018

HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD

www.heig-vd.ch

Introduction

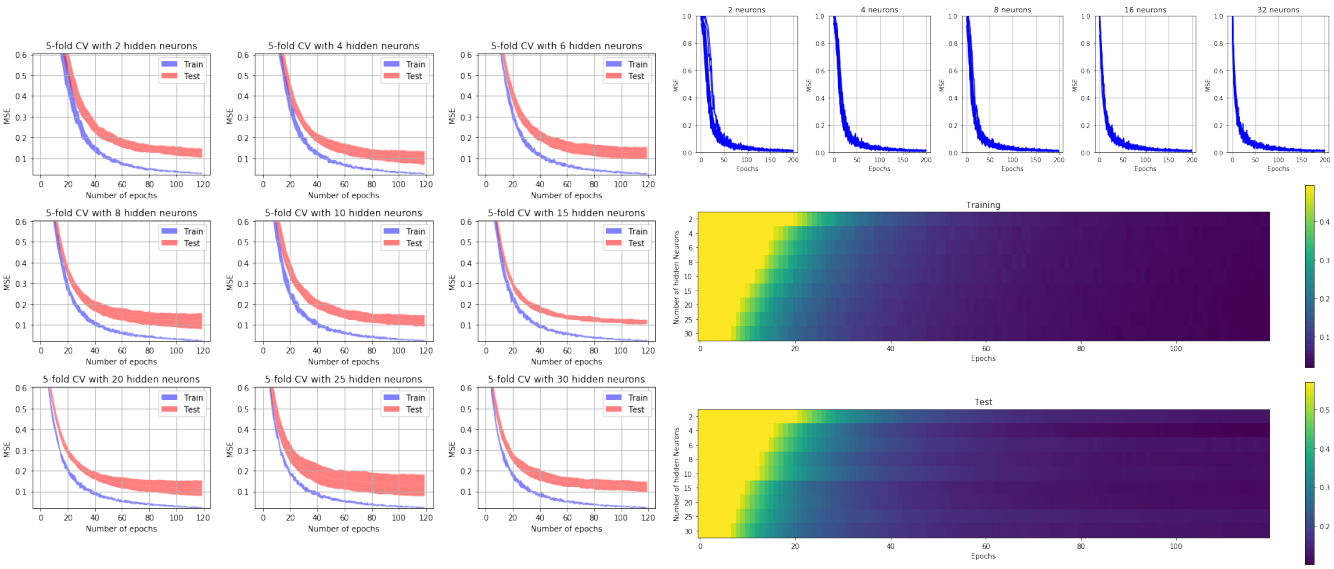
In this practical work, we trained multiple multilayer perceptrons (MLP) to classify samples of voice from men, women and children, both natural and synthetic.

Procedure

For each experiment, we trained multiple MLPs with different parameters in order to find a set of parameters yielding better results for the samples we used. Mainly, we explored the number of neurons in the hidden layer (each MLP we produced only had one) and the number of epochs used to train the model.

For the features, we start by cutting up the audio file in parts which then go through MFCC that gives us back 13 features per part. We then averaged every the list of features to obtain 13 features per file.

To visualize the results, every experiment produced three plots as follows:



The three plots express the same thing: the computed error for a given number of neurons and epochs. However, they offer a different interpretation of the results. The top right plot helps use quickly choose the number of neurons for our hidden layer and the number of epochs. We can then use the other two plots to check whether our choice made sense.

In the following pages we will only use the leftmost one as we found it easier to analyze.

Model Selection

For every model, we adjusted the hyperparameters (momentum, learning rate, etc...) until the visualization produced results we deemed sufficient. We then picked the number of neurons that yielded the best compromise in the evolution of the error for both the training and testing. Choosing the number of epochs was a matter of checking whether the plots expressed a significant reduction in the error for more epochs.

Speaker Recognition Experiments

Men vs. Women

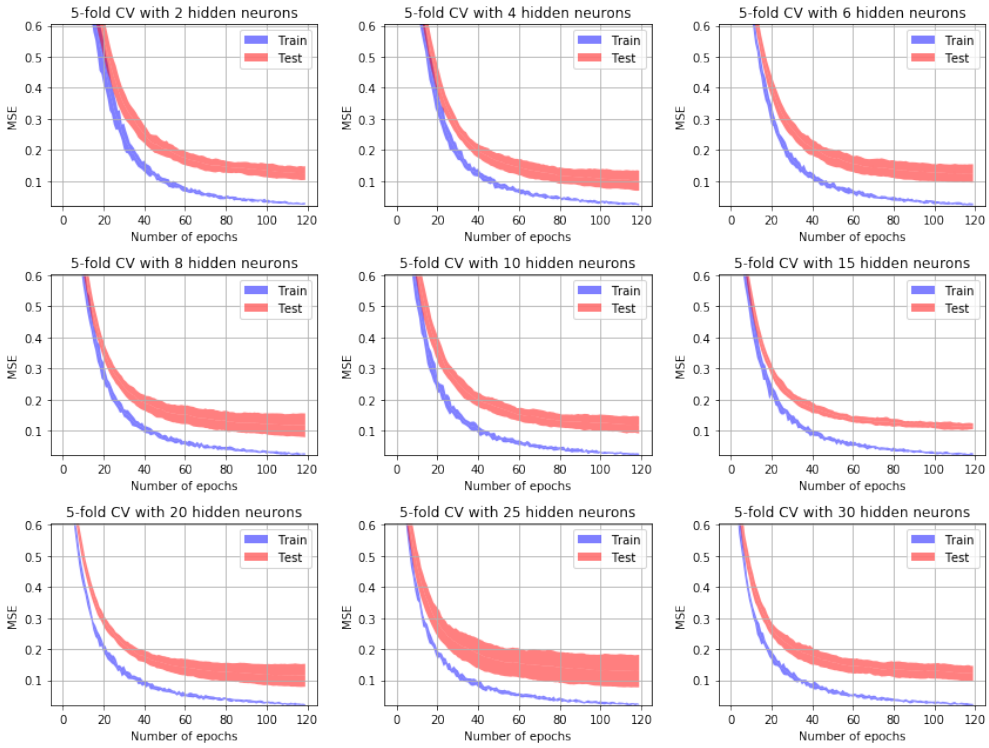


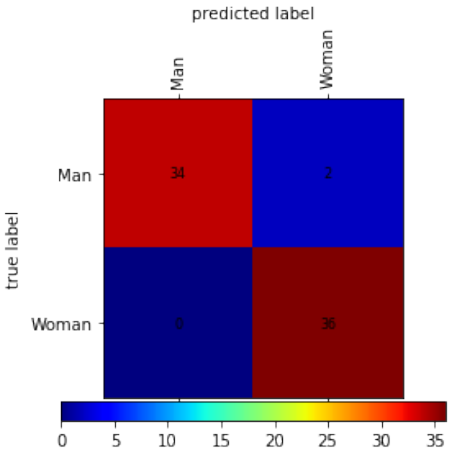
Figure 1: Testing the size of the hidden layer for Men vs. Women

Parameters

As we can observe in the plots above, using 4 neurons seems to have the "least" overfitting as the test error is closer to the train error than in the other plots. Most of the plots flatten after around 100 epochs so we deemed unnecessary to train any further. We used a learning rate of 0.001 and a momentum of 0.8. As we have only two classes to differentiate, we can use only one neurons for the output.

Results

The average F1 score of this model is 0.972. This is good as the model does not show signs of excessive overfitting.



Overfitting

We can see a bit of overfitting in the difference between the test error and the train error. The F1 score and confusion matrix also seem to indicate our model is very efficient at predicting our two classes.

Men vs. Children

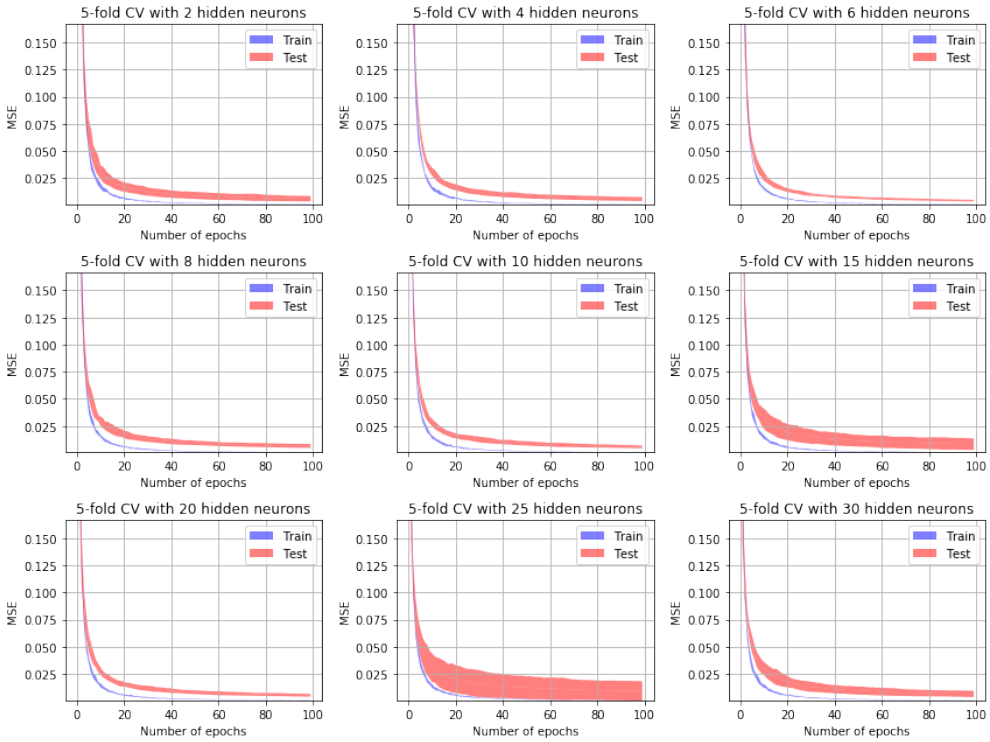


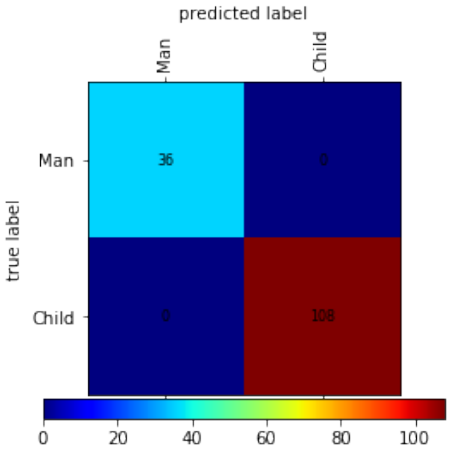
Figure 2: Testing the size of the hidden layer for Men vs. Children

Parameters

As we can see, using 6 neurons in the hidden layer seems to have a very good performance in both the training and testing phase. Because of the lowered gain after 80 epochs we decided to train until that point.

Results

The average F1 score of this model is 1. This result seems strange because of the disparity in size between our samples for the two classes (36 men samples vs 108 children samples).As the error is getting low really fast, we used a higher learning rate of 0.005 and a momentum of 0.8.



Overfitting

The results are almost too good to be true. The difference in size in the two sample sets and the very low error are good hints of overfitting. However, because the testing error is close to the training error the overfitting, if it exists, would be at the hyperparameter level.

Women vs. Children

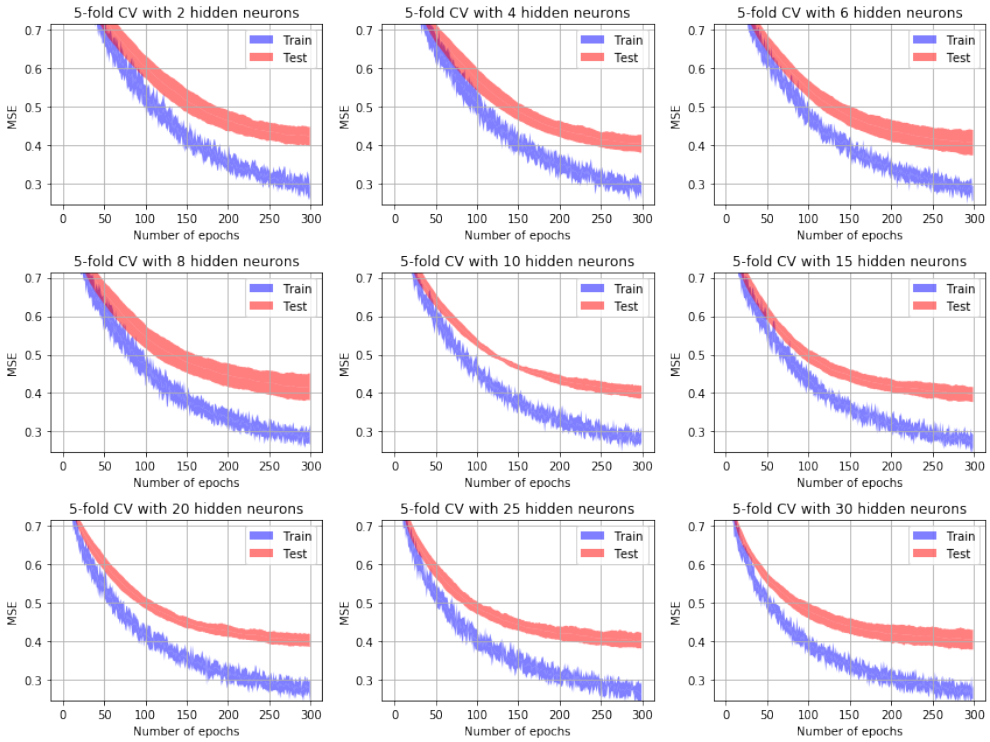


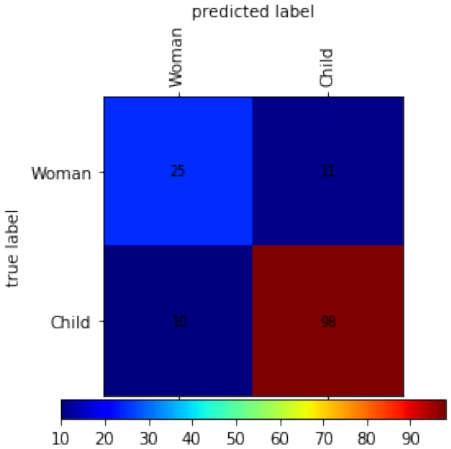
Figure 3: Testing the size of the hidden layer for Women vs. Children

Parameters

Using 10 neurons, the error is slightly closer and less spread out than the other models, so we settled for that. Nonetheless, it seems training kept reducing the error at a noticeable rate in our runs. We stopped after 300 epochs for performance reason, as running the training, testing and cross-validation on our laptops was time consuming. As the error curve oscillated a lot, we had to reduce the learning rate to 0.0005 and the momentum to 0.35.

Results

The average F1 score of this model is 0.804, with a F1 score of 0.704 for the first class and 0.903 for the second one. This, of course, is a side effect of having a lot more samples for the second class than the first one. Our model adjusted to classify what he saw more frequently. An other reason might be that the women and the children both have high-pitched voice. Therefore, it is harder to differentiate them.



Overfitting

After a hundred epochs, the error in training starts to diverge from the error in testing, which indicates our model overfits our training data. The final F1 score seems to indicate we do not suffer from it too much.

Men vs. Women vs. Children

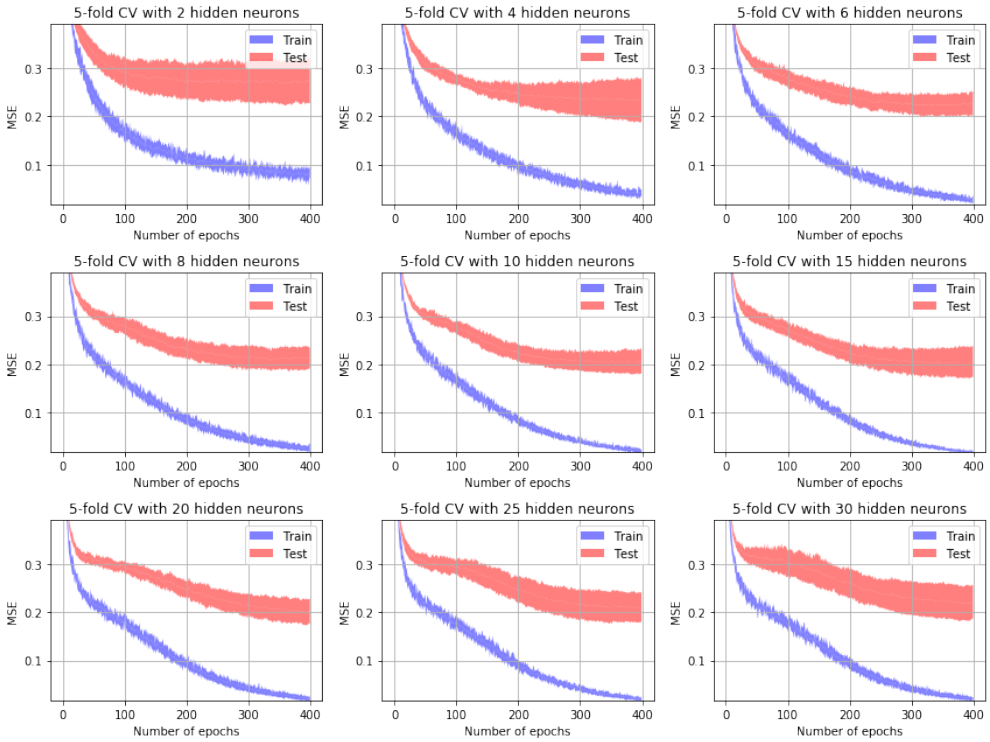


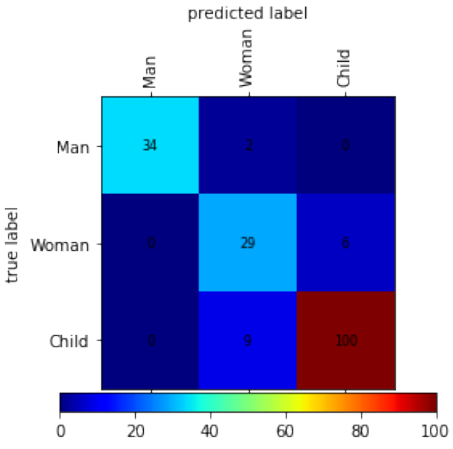
Figure 4: Testing the size of the hidden layer for Men vs. Women vs. Children

Parameters

For this experiment, we decided to pick 20 neurons for the hidden layer. Because the error kept improving, we pushed the epochs to 400 after which it became difficult to manage the experiment with our hardware. We choose a learning rate of 0.001 and a momentum of 0.08. As the cruve oscillated, we could not improve the learning rate to try to converge faster. Since we had to differentiate 3 classes, we had to use 3 neurons for the output.

Results

The average F1 score is 0.892. The second class ("Woman") has the lowest score with 0.773 whereas the two other classes have respectively 0.971 and 0.930 as their scores. We can see that it finds quite well if the input is a man, but it has more difficulty to differentiate child and woman. As said previously, the length of each dataset is not the same which influences the results. Finally we can see that we have the same problems of differentiations than when we did the experiments on each class two at a time.



Overfitting

This experiment shows a quite a bit of overfitting as the error curve in testing is clearly dissociated from the error curve in training.

Natural vs. Synthetic

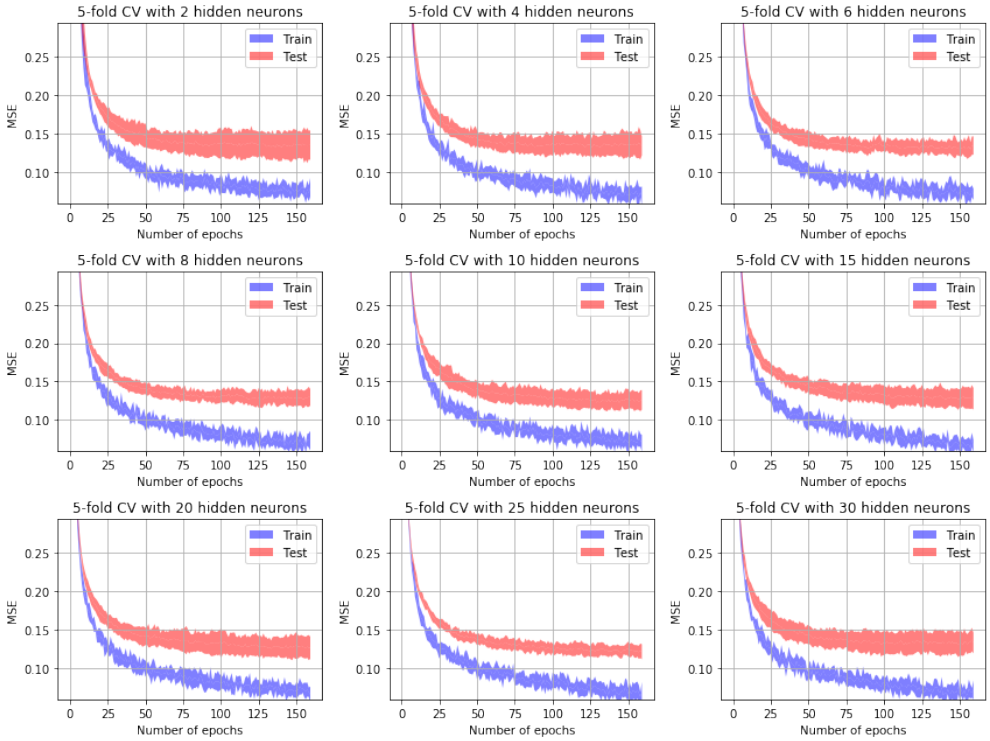


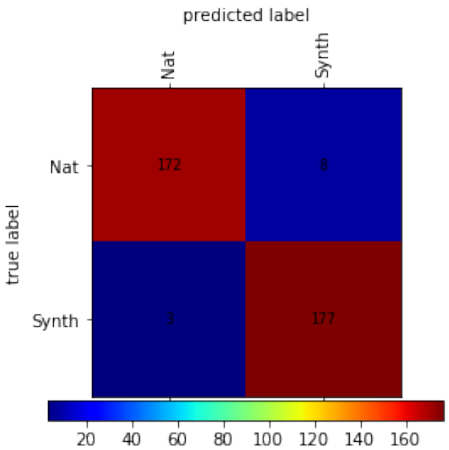
Figure 5: Testing the size of the hidden layer for Natural vs. Synthetic

Parameters

For this experiment we had a hard time choosing between 8 and 25 neurons. In practice we did not obtain wildly different results by selecting either option so we decided to have 25 neurons in the hidden layer. The number of epochs was easier of the plots stop being very steep after 120 epochs. We used a learning rate of 0.001 and a momentum of 0.8.

Results

The average F1 score is 0.969 which is very good. The overall low error rate and the fact both classes have low false positives and false negatives tells us we can properly differentiate between a synthetic and natural voice sample with our model.



Overfitting

The plots do not hint at a significant overfitting.

Women (Synth) vs. Children (Synth)

As the result of the natural women vs.natural children were not that good, we wanted to see if it was the same for the synthetic voices.

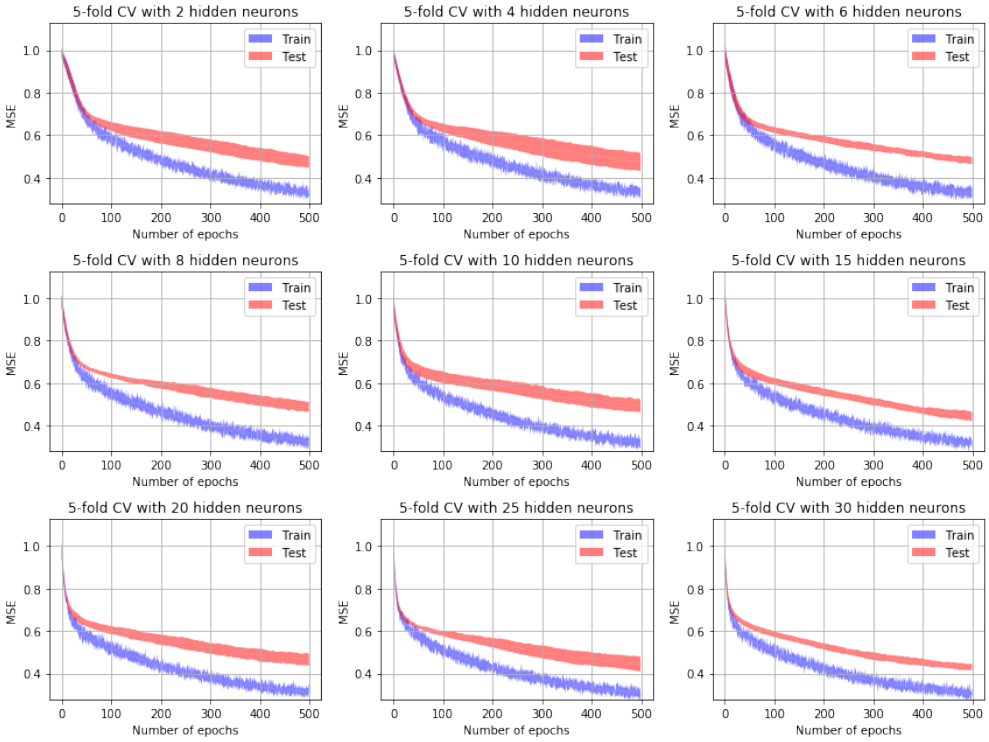


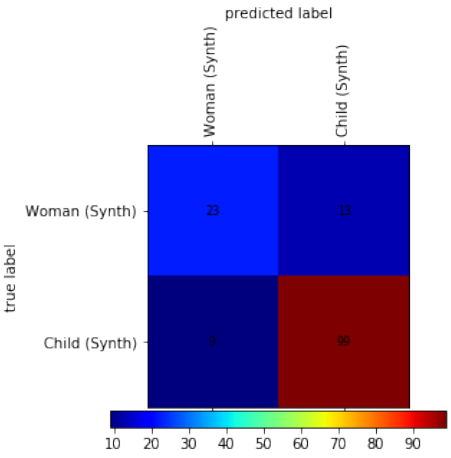
Figure 6: Testing the size of the hidden layer for Women (Synth) vs. Children (Synth)

Parameters

Overall, 30 neurons seemed to give us better results when looking at both the error curves. Since the error curves oscillated a lot, we had to reduce the learning rate and adjust the momentum. We respectively used 0.00075 and 0.35.

Results

It's interesting to see that most of the woman voices are detected as children voices. We think that it is due of the difference of size of the two datasets. The neural network is better trained to detect child voices. This explains the average F1 score of 0.788 whith respective F1 scores for the two classes being 0.676 and 0.899.



Overfitting

There seems to be a bit of overfitting if we were to infer it from the plots which is ironic when considering the average performance of the model.