MLG PW3: Speaker recognition

Guillaume Vetter, Luc Wachter

April 2020

1 Introduction

The goal of this practical work is to gain experience with neural networks, dataset manipulation and results validation. It allows us to use actual data to train a network for a classification problem. Our solution should be able to differentiate between the voices of female, male and infant speakers with reasonable accuracy.

This report contains a quick description of the problem and the data used to solve it and our account of the experiments. For each of them, we describe the parameters we chose and how we chose them and we show the performance of our final model. These steps were taken using adapted code from the provided notebook `9_model_selection`.

2 The data

The data used to train our models is comprised of 360 recordings of vowels spoken by different people (and by software). The recordings are short audio files of men, women and children pronouncing vowels. Some of these are pronounced by synthetic voices. Table 2 shows the different types of data and the number of recordings of each type.

Natural / Synthetic	**Speaker type**	**Number of recordings**
Natural	All	180
Synthetic	All	180
Natural	Female	36
Natural	Male	36
Natural	Child	108
Synthetic	Female	36
Synthetic	Male	36
Synthetic	Child	108

Table 1: Number of recordings by type

2.1 Preprocessing

As proposed by the professor, we used the Mel-Frequency Cepstral Coefficients (MFCC) to preprocess the recordings into a number of windows each summed up as 13 coefficients. We then used an aggregate function to reduce the coefficients of all the windows of each recording to only 13 values. For our three first experiments, we used the average function to do so. Thanks to this, the neural networks' number of inputs could simply be reduced to 13 (except when stated otherwise). We also tried to normalize the coefficients between -1 and 1 but the only result was slower learning speeds. We reverted to using the normal averaged coefficients after that.

2.2 Labelling

For the first two experiment, we only focus on differentiating male voices from female voices. We only need one output for this and since we use hyperbolic tan (`tanh`) as an activation function, we set male voices to equal -1 and female voices to equal 1.

For the last two experiments, we needed to differentiate between male, female and child voices. Since three outputs are necessary for this, we opted to use vectors of three elements to represent our classes. Thus, female voices were labeled $(1, -1, -1)$, male voices $(-1, 1, -1)$ and child voices $(-1, -1, 1)$.

3 First experiment

The first experiment consists of analysing natural, male and female voices. This implies using our 13 inputs for our averaged MFCCs, and a single output neuron, as explained in section 2.2.

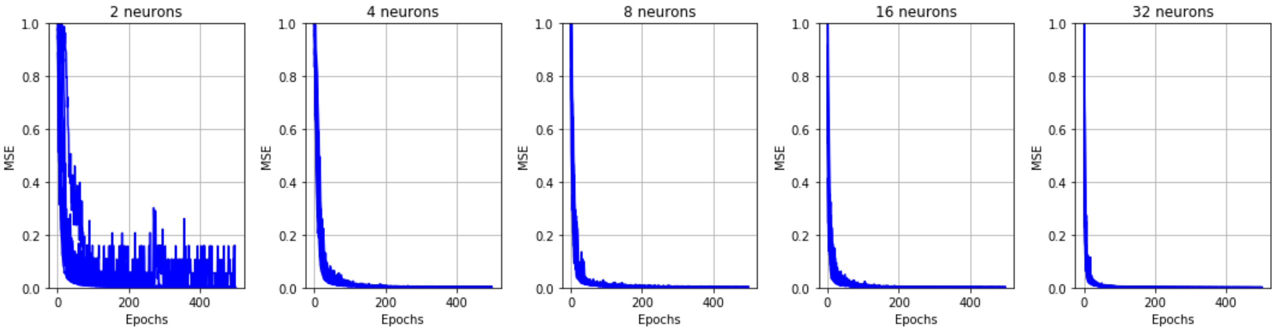
3.1 Final model parameters

As with each of the experiments, we started with the parameters in table 3.1. Those are good initial parameters and some of them were used for all the experiments (momentum, learning rate, K and the number of initialisations). Changing them did not improve the behaviour of the network in our testing, so we kept them at their sane defaults.

Parameter name	Initial value
Number of initialisations	10
Number of epochs	500
K	5
Learning rate	0.001
Momentum	0.5
Threshold	0.0

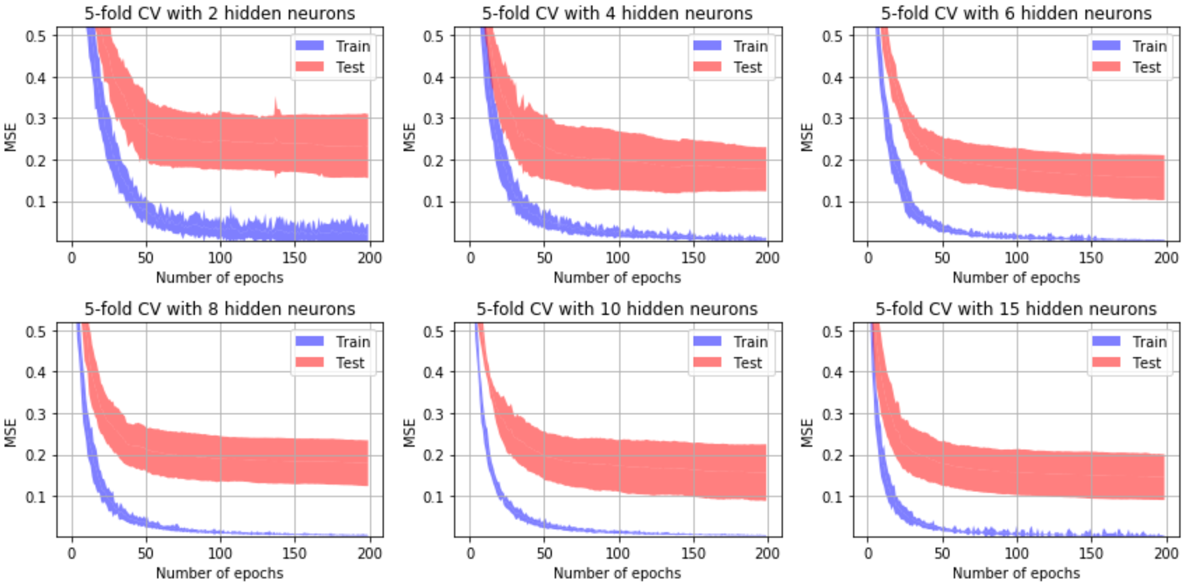
Table 2: Initial parameters used for exploring

Nevertheless, we had to create test runs to select the correct amount of epochs and for the appropriate number of hidden neurons. To do so, we generated this graph to choose how many epochs our model will use:



As we can see, the curve begins to flatten consistently after 200 epochs. We surely could have taken a bit less than that but we decided to settle with 200, because it surely is better to do more than enough than less than enough.

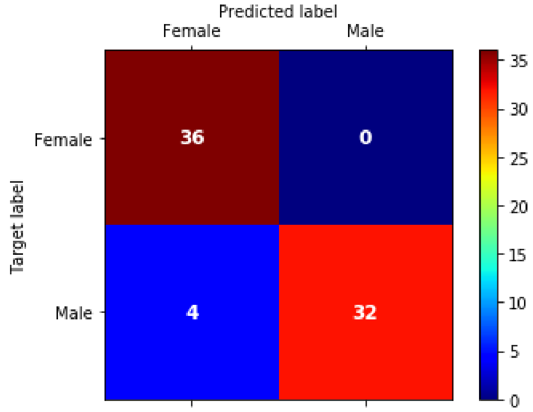
Now that we fixed the amount of epochs at 200, we selected the number of hidden neurons using these plots comparing the performance between variable models:



To choose the best one, we took the one with the least testing error (in red on the graph) and more importantly, with the smallest variation in testing error. 8 seemed to be a good choice so we decided to go with it.

3.2 Performance

The performance of our model was outstanding. The confusion matrix of the model looks like that:



As we can see, more often that not, female voices are recognised as female and male voices as male. There was no prediction error for the female voices and only 4 male samples were recognised as female. 94.4% seems like a good success rate for our model, hopefully not over-fitted. The **F1 score** is 0.941 for male voices and 0.947 for female voices).

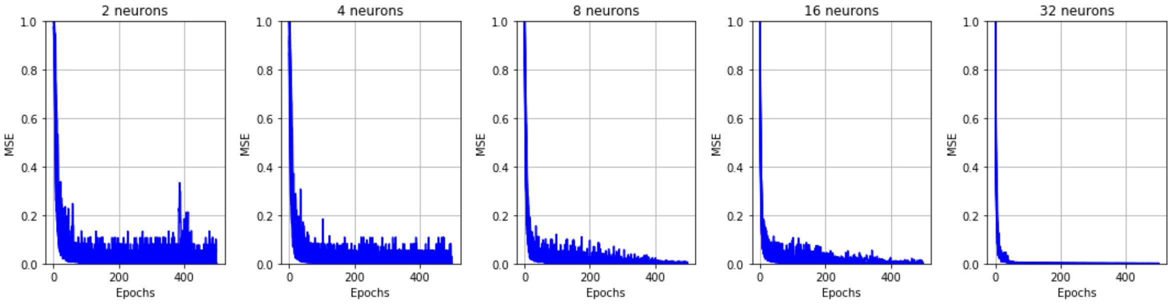
Of course, to achieve greater accuracy and to minimize the visible over-fitting, we would need more training data. We could also use data augmentation techniques or regularizers, but we are not familiar enough with these techniques yet.

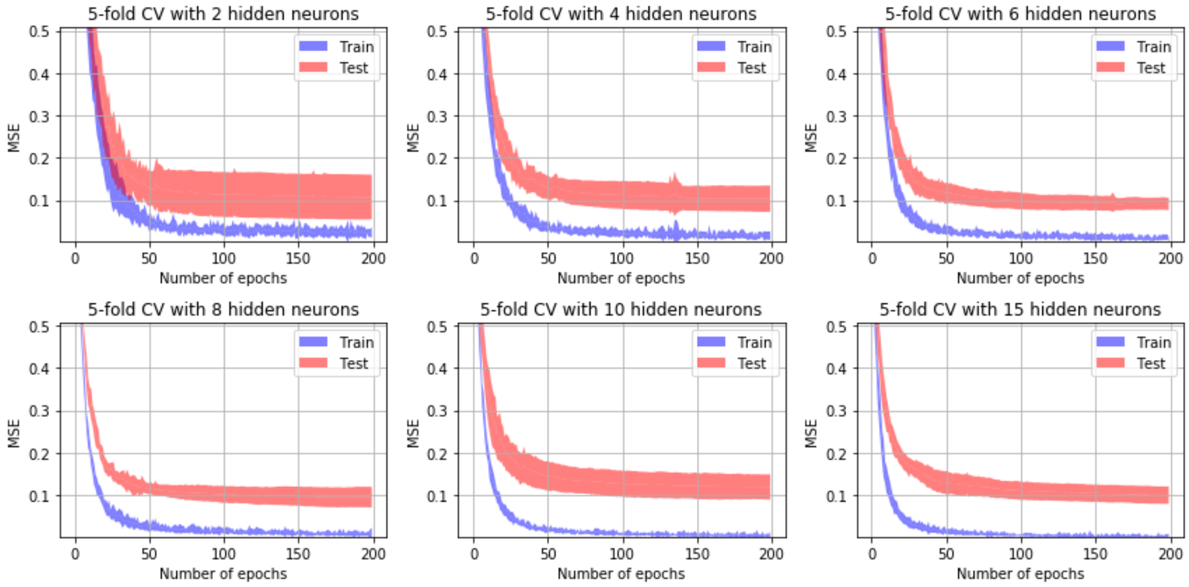
4 Second experiment

The second experiment consists of analysing natural and synthetic voices, male and female. The setup here is identical to the previous experience, since we expect the same outputs and use the same inputs.

4.1 Final model parameters

The reasoning behind the selection of the parameters is the same as in section 3. Furthermore, we chose the exact same values for our model. These are the same plots as in the first experiment:

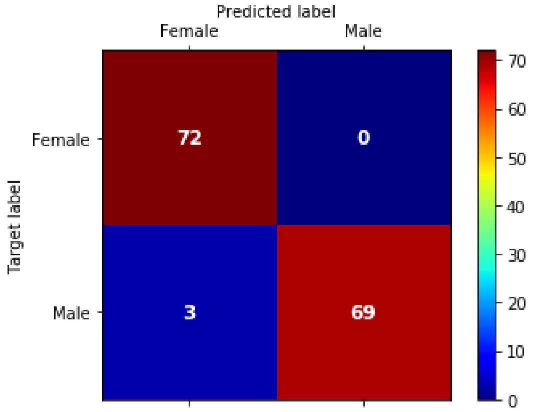




We can see that the choice of 200 epochs and 8 neurons was good for this experiment too.

4.2 Performance

The performance of the model for this experiment is even better than the one of the previous experiment. This is something we expected since doubling the data samples, even if some samples are synthetic, should help the network train better for more general cases.



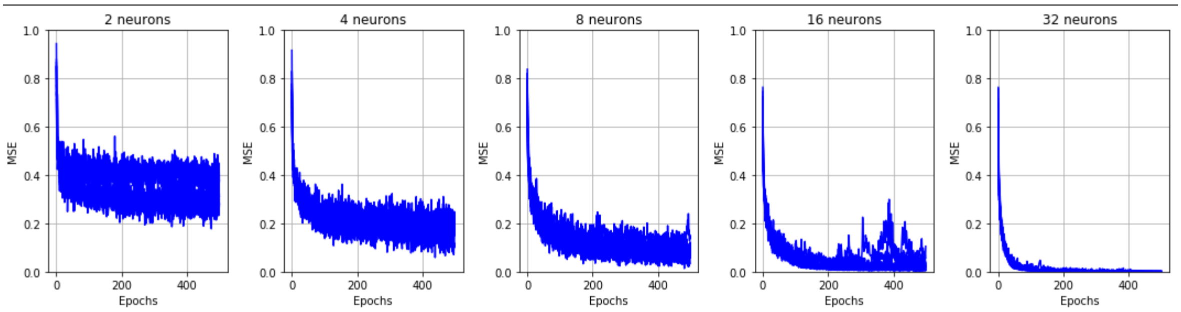
We can see that the confusion matrix is great, as there are only 3 errors over a total of 144 samples. This lead to a success rate of 97.9%. The **F1 score** for female voices is 0.979 and 0.978 for male voices.

5 Third experiment

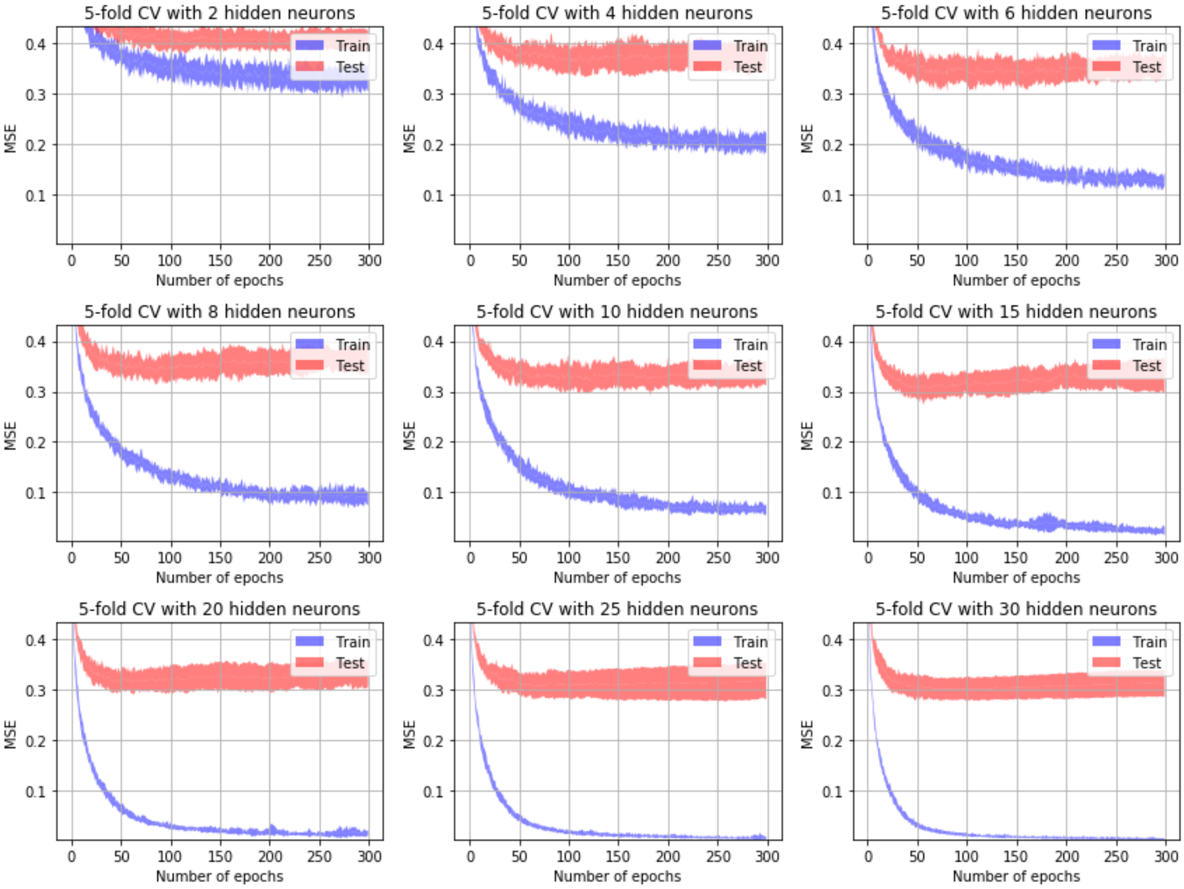
This experiment introduces children to the mix, the dataset containing now every audio files given for this assignment. Since we have three classes as outputs, we use three output neurons in the configuration described in section 2.2.

5.1 Final model parameters

For this experiment, we changed the parameters a bit. Let's look at the plots:



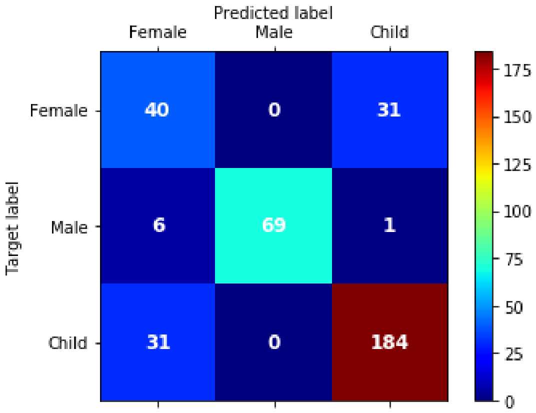
For this model, we chose 300 epochs. That seemed like an appropriate amount, since the MSE does not get any better past this point. After that, like in the previous experiments, we had to choose the number of hidden neurons:



As we can see in these graphs, more than 15 neurons seems to over-fit the model quite a lot. The training curve gets very thin, but the testing curve does not get any better. That means that the model becomes very precisely fit for our training set despite our use of cross-validation. That is not a good thing, so we decided to stick with 10 hidden neurons, since the performance seemed good for this amount. The MSE is fairly low for the training for 300 epochs: less than 0.1.

5.2 Performance

This experiment offers us a new axis of thought. As we could have expected, the model has trouble differentiating between child and female voices. It can be expected, because the pitch of a female can be very similar to the one of a kid. The confusion matrix illustrates this problem very well:



In the matrix, we can see that males are very rarely mistaken for something else, and reciprocally. Nevertheless, 43% of the female's samples are labeled as kids by the model and 14% of the kid's samples are labeled as women. The total success rate of the model is still quite good though, with a 80.9% success rate.

The **F1 score** for female voices is 0.541, 0.952 for male voices and 0.854 for child voices. This further shows that differentiating between children and women is the complicated part in this problem.

6 Fourth experiment

In this fourth experiment, we decided to add 13 new input neurons. The goal behind this is to try to improve the performance of the model in the previous experiment. We want it to have more tools to differentiate child and female voices.

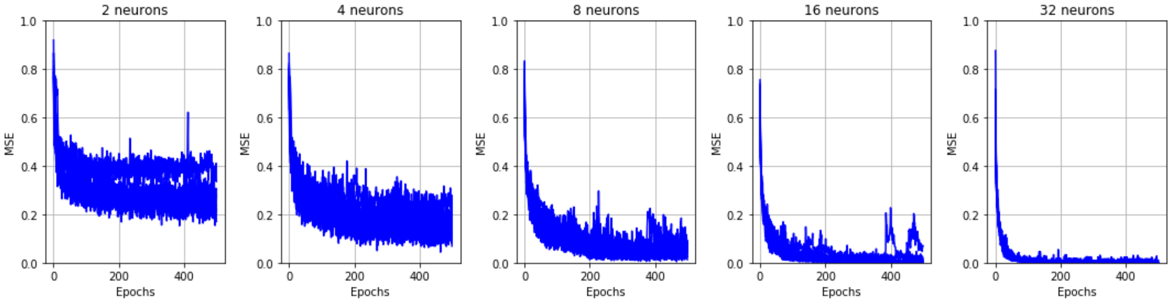
6.1 Description

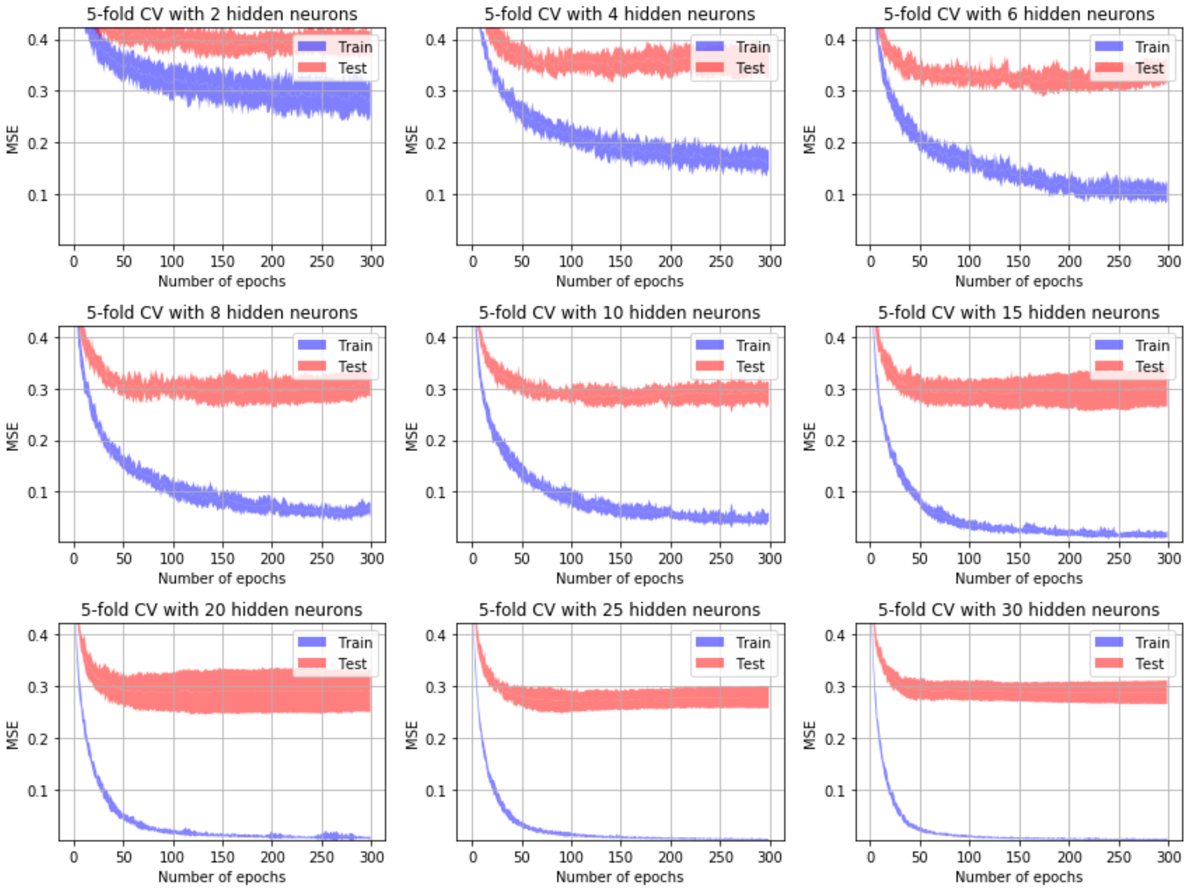
We added 13 new input neurons. The 13 original ones represented the average of every MFCC window for a recording. The 13 new ones represent the standard deviation of these window. We decided to use the standard deviation because we feel it provides new information. It is not linked to the average like the median for example. Our objective is to allow the model to see if a feature is more relevant or not based on this new input.

6.2 Final model parameters

For this final experiment, the first thing we took note of was that the running time roughly doubled. This is perfectly understandable since we have doubled the amount of input neurons. That means that we double the number of weights between the input and the hidden layer, which roughly doubles the number of computations.

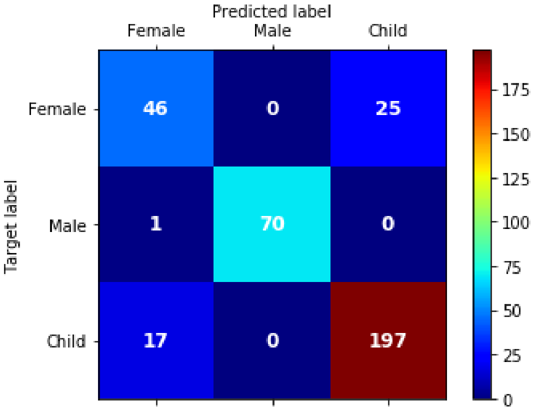
About the parameters, considering the graphs below, we made the same choices for the parameters than in experience 3, for the exact same reasons:





6.3 Performance

We can see on the following confusion matrix that there is only a slight increase in performance. Part of it is probably due to chance, but we feel there is a subtle improvement. Adding the standard deviation to the mix helped the model recognise whether or not some of the voices were from a child or a woman. Nevertheless, the improvement is minor. We almost doubled the computation time for some small percent increase. Indeed, we went from 80.9% correctness to 87.9% successful prediction with our last model.



The **F1 score** for female voices is 0.681, 0.993 for male voices and 0.904 for child voices. This once again denotes visible improvement, but we are not positive on whether the added computations are worth it.

7 Conclusion

In the end, it appears the goals of this practical work have been achieved. Our experimentation with model parameters and our interpretations of performance data has given us valuable training. Our results are satisfying, but we were still able to contrast them with observations based in pragmatism. The experiments were interesting and we feel more confident with our knowledge of neural networks now.