Audio Classifier

Evelyn Williams

1. Features

Mel-spectrograms were used as feature representations since they encode acoustic characteristics of the different classes. For example, singing spectrograms show a distinct band of intensity between the third and fourth formant (the “singer’s formant”) (Figure 1). Mel-spectrograms also encode F0 information and intensity, which other representations like MFCCs do not. For example, F0 is more likely to be sustained in singing and more variant in speaking (Figure 1). Intensity is more likely to vary over short periods in rap than in speech (Figure 1). Mel-spectrograms were extracted using a square window of length 2048 and a hop size of 512, with 80 Mel filters, then converted to the logarithmic scale then min-max scaled to the [0,1] range.

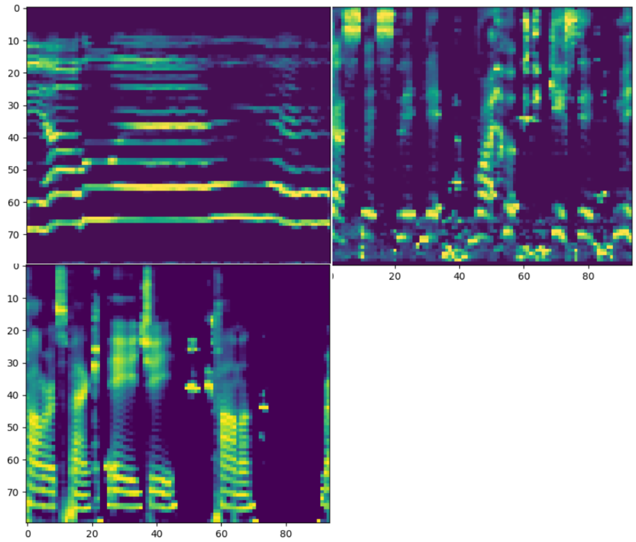

Figure 1: *Mel-spectrogram representations of singing, rap and speech, respectively.*

2. Model

I implemented a small convolutional neural network, based on the baseline model from [1] (Figure 2). I chose a convolutional model since they are good at learning image features. In this case, they should learn feature maps which represent differences in the classes’ spectral features, and in how they change over time. The model (figure 1) has two 2D convolutional layers, with a 5x5 kernel, stride=1, and dilation=1. Each convolutional layer uses a rectified linear unit activation function and is followed by a 2D max pooling layer, with a 2x2 window and stride=2, to downsize the representations by half in both dimensions. The resulting feature maps are flattened and fed through the first fully connected layer with 500 neurons, ReLU activation function, and dropout (p=0.2).

Flattening could be problematic for generalisability, since it allows the following fully connected layer to be sensitive to the location of acoustic events in the time dimension. This could, for example, decrease accuracy for audio clips with leading or trailing silences, since they make up only a small proportion of the training data. On the other hand, flattening leads to sensitivity in the frequency dimension, which could be helpful for learning acoustic features which fall into specific frequency ranges, such as the singer’s formant.

As an alternative to flattening, I also tried passing the unflattened representations through a full-connected layer with 3 neurons in the channels dimension, and taking the mean in both directions to output a 3-element vector, but this yielded much higher losses than flattening first.

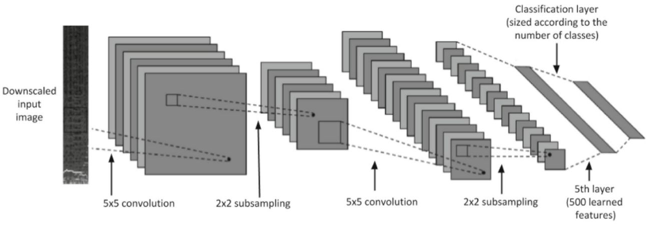

Figure 2: *Model architecture, taken from [1].*

3. Training

The dataset was randomly split into 80% training, 10% validation, and 10% test data. No effort was made to ensure the proportions of data classes were balanced in each set.

The model was trained using stochastic gradient descent. An exponential decay scheduler was used to reduce the learning rate every epoch. Hyperparameters were tuned manually by inspecting training curves at different settings and selecting the settings which produced validation losses which were both low and stable. The experimental hyperparameter values are listed in (Table 1). The model was trained for 100 epochs, but converged on the training set (loss=0.0003) by the 30th epoch (Figure 3). The final model was the one which attained the lowest loss (0.049) on the validation set (epoch 69), to reduce overfitting to the training set.

Table 1: *Hyperparameter values tested. Bold values were used in the final model.*

Hyperparameter	**Values**
Learning Rate	[0.1, **0.05**, 0.01, 0.001]
Decay Gamma	[0.9, 0.93, **0.95**, 0.97]

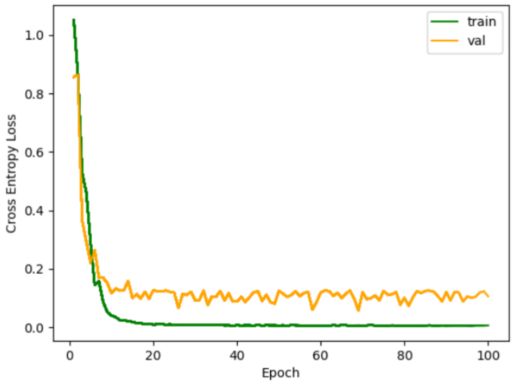

Figure 3: *Mean Cross Entropy Loss calculated on the training and validation sets.*

4. Results

The model achieved 95-96% accuracy for each class, with F1 scores between 0.93-0.97 (Table 1). Figure 4 shows the model misclassifies both speech and singing samples as rap, and vice versa, but never confuses speech and singing. This result is likely due to the spectral distinctness of speech and singing, with rap's acoustic features falling somewhere between the two. The two speech samples misclassified as rap subjectively sound more fast-paced and rhymthic than other samples, perhaps explaining this misclassification. The reasons for the misclassification of the rap and singing samples are not clear from audio-visual inspection.

Table 2: *Accuracy and F1 scores for each class.*

Class	Accuracy	F1
Speech	0.96	0.97
Rap	0.95	0.93
Singing	0.96	0.97

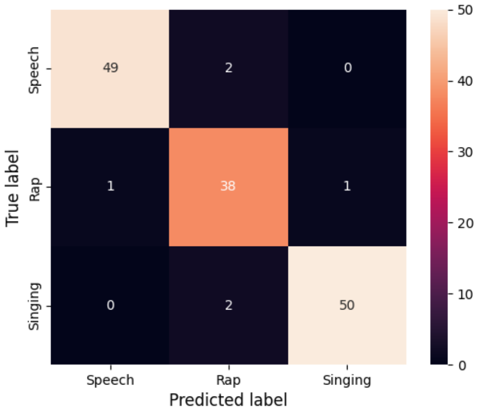

Figure 4: *Confusion matrix for the three classes.*

5. Improvements

Increasing the amount of rap examples in the training dataset may allow the network to learn a more accurate representation of this class. Balancing the proportion of each class in the training set could improve performance too, since it's possible the network simply had too few rap examples to earn its spectral features well.

Full hyperparameter tuning would likely improve prediction accuracy, but time limitations did not allow for this. A more powerful pretrained model could be more capable of learning the differences between the classes. For example, a pretrained ResNet could be fine-tuned, by reducing the hop-size in mel-spectrogram generation, and using 224 Mel filters, to output spectrograms of the required [224x224] size.

5.1. References

[1] Nanni, L., Costa, Y.M.G., Aguiar, R.L. et al. Ensemble of convolutional neural networks to improve animal audio classification. J AUDIO SPEECH MUSIC PROC. 2020, 8 (2020). https://doi.org/10.1186/s13636-020-00175-3