The development dataset was downloaded and pre-processed as follows:

1. The waveforms were converted via a short-time Fourier transform, then converted into a power spectrogram (by squaring the amplitudes), and finally converted to decibel units. (frame length: 2048, hop length: 512)
2. The spectrogram was saved as a grayscale image with sizes 64x64.
3. The mfccs (mel-frequency cepstrum coefficients) were extracted (20 mfccs).

The data were subsampled such that the number of examples for each class matched the size of the smallest class.

The data were then split into 75% training data, 5% validation data, and 20% test data.

3 different models were trained using the processed data above.

1. Model 1: 2D CNN on the grayscale images
   1. Architecture
      1. 2D Conv layer (filters: 32, kernel size: (3, 3), relu activation)
      2. 2D MaxPooling (size: (2, 2))
      3. 2D Conv layer (filters: 64, kernel size: (3, 3), relu activation)
      4. 2D MaxPooling (size: (2, 2))
      5. 2D Conv layer (filters: 64, kernel size: (3, 3), relu activation)
      6. 2D MaxPooling (size: (2, 2))
      7. Flatten
      8. Dense layer (units: 128)
      9. Dropout layer (rate: 0.5)
      10. Softmax output layer (classes: 9)
   2. Learning
      1. Adam optimiser (learning rate: 1e-3)
      2. 500 epochs (batch size: 50)
   3. Model selection: Model is evaluated every epoch on the validation set. The model with the highest accuracy is picked.
2. Model 2: 1D CNN on the mfccs
   1. Architecture
      1. 1D Conv layer (filters: 128, kernel size: 100, relu activation)
      2. 1D Conv layer (filters: 128, kernel size: 30, relu activation)
      3. 1D MaxPooling (size: 2)
      4. 1D Conv layer (filters: 128, kernel size: 15, relu activation)
      5. 1D MaxPooling (size: 2)
      6. Flatten
      7. Dense layer (units: 128)
      8. Dropout layer (rate: 0.5)
      9. Softmax output layer (classes: 9)
   2. Learning
      1. Adam optimiser (learning rate: 1e-5)
      2. 500 epochs (batch size: 50)
   3. Model selection: Model is evaluated every epoch on the validation set. The model with the highest accuracy is picked.
3. Model 3: LSTM on the mfccs
   1. Architecture
      1. Non-stateful LSTM layer (units: 64, return\_sequence: True)
      2. Non-stateful LSTM layer (units: 64, return\_sequence: True)
      3. Non-stateful LSTM layer (units: 64)
      4. Dropout layer (rate: 0.5)
      5. Softmax output layer (classes: 9)
   2. Learning
      1. Adam optimiser (learning rate: 1e-3)
      2. 500 epochs (batch size: 50)
   3. Model selection: Model is evaluated every epoch on the validation set. The model with the highest accuracy is picked.

The models are then combined. Output probabilities are put through a weighted average (Model 1: 30%, Model 2: 40%, Model 3: 30%) to obtain the final probability.