|  |
| --- |
| ELECTENG 733 Digital Signal Processing  Semester 1 2020, Department of Electrical, Computer, and Software Engineering |

**Practical Implementation Assignment 2 - Question Answer sheet**

|  |  |
| --- | --- |
| Assessment percentage: 8% of final grade | Student name: |
| Submission deadline: May 4 2020, 9 am | Student ID: |

|  |
| --- |
| **Building an Emotion recognition system from speech**  Human-Computer Interaction applications like healthcare robots, talking aids and devices that interact with humans socially require robust emotion recognition systems. Using this emotion recognition capability, Human-Computer Interactive technology can understand human emotions and respond accordingly. Emotion recongition can be done from facial expressions, speech and gestures. In this assignment, your task is to build an emotion recognition system with speech signal as the input. With the power of signal processing, statistical analysis and machine learning, robust emotion recognition systems have been developed by researchers. However, human emotions are complex, and the features of the speech signal vary depending on the age, gender, accent-type and language of the speaker. Hence, the task becomes complicated. Good knowledge about the speech signal, signal processing techniques and decision making are key stages that can help to produce better speech recognisers. In this assignment, you will go through the process of developing an emotion recogniser from speech signal, implementing majority of the concepts you learnt during Part 1 of ELECTENG 733. It will consist of a Training stage and a Testing stage.  **Instructions:**   1. You will be building the entire system going through the process of training and testing. 2. The code should be implemented in MATLAB. 3. The code should follow best practices. Refer to [[1]](https://blogs.mathworks.com/loren/2012/01/13/best-practices-for-programming-matlab/) and [[2]](https://au.mathworks.com/help/matlab/matlab_prog/matlab-code-analyzer-report.html) for this. 4. The code should be well commented, and every section should be clearly marked. Please refer to [[3]](https://www.cs.utah.edu/~germain/PPS/Topics/commenting.html) showing the best practices for comments and the format followed. The idea is that a person who does not know what you are doing should also understand the basics of the implementation. 5. You can opt to write the code using a normal MATLAB code (.m) or a Live MATLAB code (.mlx). We have used .mlx for Tutorial 3.   **Submission instructions:**  The submission will consist of:   1. Your MATLAB code (.m or .mlx files). 2. A README (notepad file) on any special instructions on how to run the code. 3. The QA sheet with some values typed into it – preferably all zipped into a folder.  * The zip file can be submitted on Canvas. * You can submit separate MATLAB codes for each Question – 1, 2, 3, 4, 5 (Even though they are variations of the same main code) * Or you can divide your code into sections for each Question. * *If the code does not run on MATLAB and produce expected results when the marker is testing it, marks will not be awarded for Questions - 1, 2, 3, 4, 5.* * The submission deadline will not be changed, and there will be 10% penalty for late submission.   From the next section onwards, details about implementing each stage of the assignment are provided. You need to implement each stage and observe the intermediate output. |
| 1. **Training stage:**   In the training stage, the block diagram given in Figure 1 can be followed. From the total database, use 80% sentences for training and the rest 20% for testing for each emotion.   |  | | --- | | D:\UOA_teaching\2020\EE733_Signal_Processing(Continuous)\For_PIA2\PIA2_Block_diagram (4).png  Figure 1: Block diagram for the training stage of the Emotion recogniser from speech |  1. **Signal accumulation:**   This stage is already done for you, as described in [[4]](https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1349.pdf). This has produced a database with emotional speech signals in it and noise removal has been done. A part of this database is given to you as Emotional\_speech\_database.zip. This contains speech signals from *4 emotions – Angry, Happy, Sad, Excited*, from *one male and one female speaker of New Zealand English.*   1. **Feature extraction:**   This stage has to be done for each emotion set *Angry, Happy, Sad, Excited* separately.   1. **Converting speech to frames –** The speech signal needs to be read into MATLAB. The sampling frequency for all the waveforms if 44100Hz. These speech signals should be converted to frames of size 20ms for analysis. (The reasons for this have been discussed during lectures and tutorials).   For example:  If the overall length of a speech signal is 2 seconds, and the sampling frequency is 44100 Hz= 44100 samples/second, then the total number of samples will be:  44100 samples/second 2 seconds = 88200 samples.  If 20 ms analysis frame is considered, then each frame will have 44100 samples/second ms = 88 samples.  If the total length of the speech signal is 2s, then it will consist of 2s/20ms = 1000 frames of 20 ms length.   1. **Feature extraction -** Feature extraction has to be done for each frame. *Do not use inbuilt MATLAB functions for this extrcation. Use the methods discussed during the lectures and tutorials.* There are 4 features to be extracted as a requirement for this assignment.  * *Short-time energy:* This is the energy measure for 1 frame. It has to be convered to magnitude in dB for better representation. * *Zero-Crossing rate:* This is the number of zero crossings per second for 1 frame. * *Pitch:* This is the Pitch in Hz for 1 frame. To measure the pitch of a frame, first the frame as to be classified as voiced/unvoiced depending on the short-term energy and Zero crossing rate. If the frame is voiced, then its pitch can be estimated using autocorrelation. If the frame is unvoiced, then no pitch estimation is needed. * *Spectral energy:* This is the magnitude of the Fourier transform of the speech frame. The spectral energy should be in dB.   Once the feature for all the frames of a speech signal are computed, then an average across all frames for that speech signal is taken – which is the *expectation calculation*. At the end of one iteration of this process, you will have a single feature value for each of the features for one speech signal. For example (this is only an example, not exact values), for speech signal male2\_angry\_1a\_2:  Short-time energy in dB = 10 dB  ZCR =100 zero crossings/second  Pitch = 130 Hz  Spectral energy in dB =20 dB  Such values for each speech signal can be obtained. This has to be done for all speech signals corresponding to one emotion, to prepare data to build the probability model for that emotion.   1. **Model Training –** Here the probability distribution model for each feature for a particular emotion has to be estimated. Maximum Likelihood estimation can be used for this. Also, the prior probabilities have to be obtained – this could be using the classical approach or relative frequency approach. This will give all parameters needed for the Naïve Bayes classifier for one emotion based on all the 4 features.   This model training has to be done for each of the emotions separately to develop probability distribution models for *each feature for each emotion*. |
| 1. **Testing**   For testing, the block diagram given in Figure 2 can be used. From the total database, use 80% sentences for training and the rest 20% for testing for each emotion.   |  | | --- | | D:\UOA_teaching\2020\EE733_Signal_Processing(Continuous)\For_PIA2\PIA2_testing.pngFigure 2: Testing stages for emotion recogniser using speech |   In the testing stage, the signal that is tested has to go through the same Signal accumulation and Feature extraction as the training stage. Only then the results can be an indication of the system performance.   1. **Signal accumulation –** This step is already done for you, and test sentences are recorded and noise removed. From the total database, use 80% sentences for training and the rest 20% for testing. 2. **Feature extraction –** Extract the same features as the training process, using the same extraction techniques – like converting speech signal to frames, finding the feature value for each frame and then calculating its expectation for the speech signal. 3. **Model testing** – In the testing stage, we need to check the match of the new signal features to each of the emotion models. For this, calculate the likelihoods for each emotion, for all features. Then combine the likelihood together to make the Naïve Bayes classifier decision. The emotion tag should be your output. (We have discussed in detail about how to calculate likelihoods, and combine multiple likelihoods together, and make the Naïve Bayes classifier decision during lectures and tutorials.) |
| 1. **Performance analysis**   Once a classification system is built, it is essential to analyse the performance of the system. For performance analysis, we will use only the hit rate as the measure.  You are using 20% of the database for each emotion for testing. Perform the testing (Section II) the 20% of the sentences, and check if they are correctly classified to the emotion category marked on its file name. Count the number of correct classification (hits) and divide by the total number of speech signals checked to calculate the hit rate %. |
| **Question 1a:**  Implement a MATLAB code performing training and testing as mentioned above for the male2 database alone.  **Question 1b:**  Conduct performance analysis for each of the 4 emotions and combined for all emotions for male2, and visualise this information using a barplot (similar to barplot discussed in Tutorial 3).   |  | | --- | | Insert Barplot for Question 1b here |   **Question 2a:**  Implement a MATLAB code performing training and testing as mentioned above for the female1 database alone.  **Question 2b:**  Conduct performance analysis for each of the 4 emotions and combined for all emotions for female1, and visualise this information using a barplot.   |  | | --- | | Insert Barplot for Question 2b here |   **Question 3a:**  Implement a MATLAB code performing training and testing as mentioned above combining male2 anf female2 databases together as a single database.  **Question 3b:**  Conduct performance analysis for each of the 4 emotions and combined for all emotions for the database in 3a, and visualise this information using a barplot.   |  | | --- | | Insert Barplot for Question 3b here |   **Question 4a:**  Implement a MATLAB code performing training and testing as mentioned above for the male2 database alone for features Short time energy and Pitch.  **Question 4b:**  Conduct performance analysis for each of the 4 emotions and combined for all emotions for Question 4a, and visualise this information using a barplot.   |  | | --- | | Insert Barplot for Question 4b here |   **Question 5a:**  Implement a MATLAB code performing training and testing as mentioned above for the male2 database alone for features Spectral energy and Zero Crossing Rate.  **Question 5b:**  Conduct performance analysis for each of the 4 emotions and combined for all emotions for Question 5a, and visualise this information using a barplot.   |  | | --- | | Insert Barplot for Question 5b here |   **Question 6: Summarise**   1. Which of the following training sets had better performance (overall for all emotions)? (Tick one) 2. male2 alone 3. female1 alone 4. combined database 5. Which of the following feature sets had better performance (overall for all emotions)? (Tick one) 6. Spectral energy + ZCR 7. Short-time energy + pitch 8. Combining all 4 features 9. For male2 speaker – which emotion had the highest hit rate? (Tick one) 10. Angry 11. Happy 12. Sad 13. Excited 14. For female1 speaker – which emotion had the highest hit rate? (Tick one) 15. Angry 16. Happy 17. Sad 18. Excited 19. For the combined database, which emotion has the highest hit rate? (Tick one) 20. Angry 21. Happy 22. Sad 23. Excited 24. Why is the combined database hit rate different from individual male2 and female1 hit rates?  |  | | --- | |  |  1. Does adding more features help improve the performance of the emotion classifier?  |  | | --- | |  |  1. Provide details of the likelihood estimation after fitting a normal distribution to each feature.  |  |  |  |  |  | | --- | --- | --- | --- | --- | | Speaker | Emotion | Feature | Mean | Standard deviation | | male2 | angry | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | happy | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | sad | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | excited | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | female1 | angry | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | happy | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | sad | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | excited | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | male2+female1 | angry | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | happy | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | sad | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  | | excited | Short-time energy |  |  | | Pitch |  |  | | ZCR |  |  | | Spectral energy |  |  |   **Ungraded questions**  What did you learn from this assignment?   |  | | --- | |  |   Now, can you build a speech signal feature extraction and classification system by yourself?   |  | | --- | |  |   Did you enjoy this assignment? (Tick one)   |  |  | | --- | --- | |  |  | |

**References:**

[1] <https://blogs.mathworks.com/loren/2012/01/13/best-practices-for-programming-matlab/>

[2] <https://au.mathworks.com/help/matlab/matlab_prog/matlab-code-analyzer-report.html>

[3] <https://www.cs.utah.edu/~germain/PPS/Topics/commenting.html>

[4] [paper] J James, L Tian, CI Watson, An Open Source Emotional Speech Corpus for Human Robot Interaction Applications, Interspeech, 2768-2772