A logo of a deer and sword

Description automatically generated

**SPEECH RECOGNITION GROUP COURSEWORK – COURSEWORK 2 - EEEM030**

**Authors: Xiaoguang Liang, Tofarati Onatunde, and Zeyad Adelmonem**

**ABSTRACT**

**I WILL WRITE THE ABSTACT AS SOON AS EVERYONE HAS PUT IN THEIR WORK TONIGHT - FROM TOFARATI ONATUNDE**

Table 1: Contribution list of the written report

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Section | Abstract | Introduction | 1 | 2 | 3 | 4 | 5 | Conclusion |
| Name | Tofarati Onatunde | Tofarati Onatunde | Tofarati Onatunde | Xiaoguang Liang | Zeyad Adelmonem | Xiaoguang Liang | Zeyad Adelmonem |  |
| URN | 6632165 | 6632165 | 6632165 |  |  |  |  |  |

**TABLE OF CONTENTS**

Introduction…………………………………………………………………………………………………….4

1) Extract MFCC acoustic features……………………………………………………………………………..4

1.1) Significance of Feature Extraction in Speech Recognition………………………………………..4  
 1.2) Audio Data Setup………………………………………………………………………………….4  
 1.3) MFCC Extraction Process………………………………………………………………………5  
 1.4) Visualisation of Results………………………………………………………………………….6  
 1.5) Applications and Implications…………………………………………………………………….7

2) Initialise a set of prototype HMMs for each word in the vocabulary

3) Train the HMMs with the training data

4) Evaluate the recognizer at the at each iteration on the development data

5) Evaluate the recognizer on the supplied test data in the evaluation set

Conclusion

References

Appendix A – Code for feature extraction………………………………………………………………….

Appendix B – Main Code……………………………………………………………………………………..

**LIST OF FIGURES**

*Figure 1: Audio Signals of sound samples from the Developmet data set // 1a) ‘sp15a\_w02\_heed.mp3’, 1b) ‘sp01a\_w02\_hod.mp3’, and 1c) ‘sp30a\_w02\_heard.mp3’*

*Figure 2: Audio Waveform of ‘sp01a\_w01\_heed.mp3’*

*Figure 3: MFCC spectogram of ‘sp01a\_w01\_heed.mp3’*

*Figure 4: MFCC spectogram of ‘sp01a\_w01\_heed.mp3’*

*Figure 5: Evolution of the first MFCC coefficent*

*Figure 6: Histogram of ‘sp01a\_w01\_heed.mp3’*

*Figure 7: 3D Plot of ‘sp01a\_w01\_heed.mp3’ MFCC Features*

**INTRODUCTION**

This assignment involves designing and implementing a basic speech recognition system using fundamental speech and audio processing techniques. The system, developed in MATLAB, performs isolated-word recognition for a vocabulary of eleven words, structured around five key tasks: feature extraction, model initialisation, training, evaluation, and testing. Hidden Markov Models (HMMs) form the core methodology, providing a framework to explore the impact of maximum likelihood training on system performance. The recogniser uses the Baum-Welch algorithm and the Viterbi algorithm for decoding and recognising these words. The activities include within this coursework include:

* **MFCC Feature Extraction** for extracting acoustic features for training, evaluation, and additional test datasets.
* **HMM Initialisation and Training:** For initialising the prototype HMMs using global statistics of the training data. In addition to iteratively refining model parameters using forward-backward algorithms to compute likelihoods and re-estimate transitions and emissions.
* **Recognition and Evaluation:** for calculating the Viterbi likelihoods and decoding the recognition outputs. In turn scoring the system performance via confusion matrices and error rate calculations.

**1) FEATURE EXTRACTION**

**1.1) Significance of Feature Extraction in Speech Recognition**

Feature extraction is an essential step in developing a speech recogniser system, as it transforms raw audio signals into a meaningful and compact interpretations that can be effectively processed by the speech recognition model

Therefore, this section of the report will provide an overview of the underlying principles, processes and parameters used to extract 13 Mel-frequency cepstral coefficients (MFCCs) including the zeroth coefficient, from the development and evaluation audio files sets, while supplementing relevant images of key results.

Before, explaining how the 13 MFCCs were extracted, it is vital to note that MFCC’s are widely regarded as the standard interpretation for speech recognition. This is because MFCC’s closely mimic the way the human ear process sound, by capturing the frequency spectrum of the audio signals on the Mel scale that closely aligns with the human auditory sensitivity to sound frequencies, while actively discarding irrelevant information such as pitch variations. Which works to retain essential phonetic features needed for distinguishing between speech sounds.

**1.2) Audio Data Setup**

A graph of a sound wave

Description automatically generatedA graph of a sound wave

Description automatically generated A graph of a sound wave

Description automatically generated

*Figure 1: Audio Signals of sound samples from the Developmet data set   
 1a) ‘sp15a\_w02\_heed.mp3’, 1b) ‘sp01a\_w02\_hod.mp3’, and 1c) ‘sp30a\_w02\_heard.mp3’*

As shown in Appendix B, the MATLAB code begins by defining 30 speakers used in the development data set and listing their repeated eleven-word vocabulary collections: ‘heed, hid, head, had, hard, hud, hod, hoard, hood, whod, and heard’. Resulting in the 330 audio files showcasing variations in vowel sounds, constants and language articulations, ideal for studying phonetic features, visually represented by figure 1.

Note: As shown in appendix B, two ‘for’ loops were used which generated filenames for all speaker-word combinations. Resulting in paths to all audio files, ensuring an organised dataset for subsequent processing. Each file followed the naming convention spXXa\_wYY\_word.mp3, where: ‘XX’ is the speaker index, ‘YY’ is the word index, and ‘word’ is the spoken terminology.

**1.3) MFCC Extraction Process**

Therefore, to initialise the feature extraction of the 13 MFCCs, including the zeroth coefficient, from the development set, and later the evaluation set, a few steps had to be taken, and are described below.

**1.3.1) Frame and Hop Size Parameters**

The process begins with segmenting the audio signals into overlapping frames, each 30ms in duration. This temporal resolution corresponds to the typical length of phonemes, ensuring that each frame captures meaningful variations in speech without losing important details. Additionally, a hop size of 10ms with a 20ms overlap between consecutive frames was implemented. As this provided smooth transitions and comprehensive temporal coverage between frames. Overlapping frames are crucial in speech processing as they capture the continuity of spoken words, avoiding information loss between frames.

**1.3.2) Hamming Window for Spectral Analysis**

Noye: A Hamming window was also applied to each frame to reduce spectral leakage caused by transitions abrupt frame boundaries. This step was utilised to improve the accuracy of frequency domain analysis, ensuring that the extracted features are reliable and precise.

**1.3.3) MFCC Calculation**

Finally, a built-in MATLAB function was implemented to extract the final MFCC values from each frame. Providing a compact yet detailed representation of the speech signal that focusing on the spectral content most relevant to human perception. It utilised the previously specified frame size, hop size, and the Hamming window to ensure high-quality feature extraction. The resulting 13-dimensional feature vector for each frame captures the static spectral characteristics of the speech signal. Then process was repeated for the evaluation set of audio samples.

Take note that delta and delta-delta features (velocities and accelerations) were excluded from the overall calculation of the MFCC’s. This simplification allows the code to focus on the static spectral characteristics of the audio signals. Which in turn improves computational efficiency while also maintaining sufficient information for effective speech recognition.

**1.4) Visualisation of Results**

After extracting the MFCC features for development and evaluation audio sets, several visualisations were created to illustrate the properties of the extracted features. Note: These graphs shown in this subsection were produced for all audio samples. However, only images pertaining to audio *‘sp01a\_w01\_heed.mp3’* will be shown for simplication purposes.

**1.4.1) Audio Waveform to MFCC Spectrogram**

A graph of a waveform

Description automatically generated  
*Figure 2: Audio Waveform of ‘sp01a\_w01\_heed.mp3’*

By sourcing the length of the audio signal and plotting it against its linear amplitude, a graph providinga time-domain representation of the heed speech signal showcasing its temporal variations and intensity was made.

A yellow square with a blue and yellow square

Description automatically generated with medium confidence  
*Figure 3: MFCC spectogram of ‘sp01a\_w01\_heed.mp3’*

After putting the original audio signal, through the MFCC process described in section 1.3, the corresponding MFCCs can be displayed in a spectrogram-like image. Where the x-axis represents the frame index (time), and the y-axis illustrates the MFCC coefficients, and the colour intensity indicates the magnitude. This visualisation provides a view of how the audios spectral content evolves over time.

**1.4.2) Spectrogram of the Audio Signal**

A close-up of a graph

Description automatically generated *Figure 4: MFCC spectogram of ‘sp01a\_w01\_heed.mp3’*

Additionally, another spectrogram can be made to visualises the frequency domain of the signal over time. In turn, offering insights into the full frequency spectrum of the audio data.

**1.4.4) Evolution of MFCC Coefficients**

A graph with a line

Description automatically generated*Figure 5: Evolution of the first MFCC coefficent*

Moreover, the evolution of the first MFCC coefficient can be plotted over time to highlight its dynamic behaviour across frames. This helps to emphasis the temporal variations in spectral content, which can be utilised when distinguishing between different speech sounds.

**1.4.3) Histogram of MFCC Coefficients**

A graph with blue bars

Description automatically generated *Figure 6: Histogram of ‘sp01a\_w01\_heed.mp3’ MFCC coefficents*

The histogram of the first MFCC coefficient shows its distribution across frames. This statistical representation reveals the range and frequency of phonetic feature values, thus providing insight into the acoustic properties of each speech signal.

**1.4.4) 3D Plot of MFCC Features**

A graph of lines and lines

Description automatically generated with medium confidence*Figure 7: 3D Plot of ‘sp01a\_w01\_heed.mp3’ MFCC Features*

Lastly a 3D mesh plot can be used to visualise the first three MFCC coefficients across time. Thus, illustrating the relationships between coefficients and their temporal evolution, offering a more comprehensive view of the feature space.

These images are useful in analysing and refining the extracted features, because: (a) The MFCC spectrogram highlights phonetic variations. (b) The waveform and audio spectrogram offer complementary views of the signal’s temporal and frequency characteristics. (c) The coefficient evolution and histograms reveal trends and statistical properties, aiding in fine-tuning recognition models.

In turn allowing us to understand MFCC features and how they evolve over time. They provide valuable insights for further model development, enabling applications such as automatic speech recognition (ASR), speaker identification, and audio classification.

To conclude this section,feature extraction detailed lays the foundation for an effective speech recognition system. Demonstrating the power of how extracting the MFFC’s creates a well-suited starting point for capturing meaningful information from speech signals and training Hidden Markov Models (HMMs), of which are detailed in later sections of this report.

**END OF SECTION 1 – START SECTION 2 NEXT**

**Conclusion**

Overall, this coursework bridges theoretical concepts and practical application. Highlighting the strengths and limitations of HMM-based speech recognition on provided datasets and additional user-recorded data. By analysing the results, the project offers insights into the effectiveness of maximum likelihood training and its role in improving recognition accuracy.

**Appendix A – Code for feature extraction**

A screenshot of a computer screen

Description automatically generated A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Appendix B – Main Code (LIANG PUT YOUR CODE HERE)**