School of Computing Sciences

# COURSEWORK ASSIGNMENT

| | |
|---|---|
| **MODULE:** | CMP-6026A– Audio-visual processing |
| **ASSIGNMENT TITLE:** | Design, implementation and evaluation of a speech recognition system |
| **DATE SET:** | Week 3 |
| **PRACTICAL DEMONSTRATION:** | Week 7 Wednesday (slot to be advised) – 10 November 2021 |
| **WRITTEN SUBMISSION:** | Monday of Week 8 – 15 November 2021 |
| **RETURN DATE:** | Friday of Week 9 – 26 November 2021 |
| **ASSIGNMENT VALUE:** | 50% |
| **SET BY:** | BPM |
| **CHECKED BY:** | DG |

## LEARNING OUTCOMES

- Explain how humans produce speech from audio and visual perspectives and how these differ across different speech sounds and be able to give examples of how these are subject to noise and distortion
- Apply a range of tools to display and process audio and visual signals and be able to analyse these to find structure and identify sound or visual events
- Transfer knowledge learnt into code that extracts useful features from audio and visual data to provide robust and discriminative information in a compact format and apply this to machine learning methods
- Design and construct audio and visual speech recognisers and evaluate their performance under varying adverse operating conditions
- Work in a small team and organise work appropriately using simple project management techniques before demonstrating accomplishments within a professional setting

## SPECIFICATION

### Overview

This assignment involves the design, implementation and evaluation of a speaker-dependent speech recognition system to recognise the names of 20 students taken from the CMP-6026A/CMP-7016A modules in clean and noisy conditions. **The work is to be undertaken in pairs**.

### Description

The task of building and testing a speech recogniser can be broken down into five stages:

i) Speech data collection and labelling
ii) Feature extraction
iii) Acoustic modelling
iv) Noise compensation
v) Evaluation

The speech recogniser is to be speaker-dependent which means that it is trained on speech from just a single speaker and should also be tested on speech from just that speaker. The vocabulary is a set of 20 names taken from the students studying the CMP-6026A and CMP-7016A modules:

**Albi, Alex, Alexander, Alejandro, Aurelie, Benjamin, Brennan, Felipe, Harry, Hemal, Hugo, Max, Nathaniel, Owen, Ruby, Ruaridh, Sarah, Sophie, Vav, Yan**

The twenty names have been selected to give some words that are distinctive, some that are confusable with others and some that are short.

The assignment is to be carried out in pairs with 35% of marks awarded as a group mark with the remaining 65% as an individual mark (see below for details of the marking scheme).

The assignment will use a combination of MATLAB (for implementing feature extraction and noise compensation) and commercial tools such as SFS/WaveSurfer and HTK. These are standard tools and give a good introduction as to how such a task may be carried out in industry.

The second assignment (CW2) will be based closely on this assignment. This means that this assignment will form an important underpinning for the next coursework. Feedback and feedforward from this assignment should be useful when undertaking the second assignment.

## 1. Speech data collection and labelling

A speech recogniser must be trained on examples of the speech sounds that it is expected to recognise. For this assignment the vocabulary of the speech recogniser comprises 20 names taken from students on CMP-6026A and CMP-7016A. Therefore, the first part of the assignment involves recording examples of each name in the vocabulary. Theoretically, the more examples of each name, the higher the accuracy of the speech recogniser. A suitable number of examples to start with is 10 or 20 of each name. Each speech file can be stored as a WAV file (for example `speech001.wav`).

Next, an associated label file (`.lab` file) needs to be created for each audio file. This should contain the start and end times of each word and for instances of silence in the WAV file. This follows on from what was looked at in Lab Sheet 1. A typical label file (for example `speech001.lab`) will take the form:

```
0 3200000 silence
3200000 6600000 Alex
6600000 7800000 silence
```

In HTK, time is represented in 100ns units, which gives in the large numbers seen above.

For ease of file management, the WAV file and label file should use the same core filename, but with different extensions.

It may be easier to record all 20 names in one WAV file rather than an individual file for each name.

## 2. Feature extraction

Feature extraction's task is to extract a set of feature vectors from each speech utterance that forms the input to the speech recogniser. This involves designing and implementing in MATLAB an algorithm to extract feature vectors from the speech waveform. Many different feature extraction methods exist, but for this assignment you should consider only filterbank-derived cepstral features. You may first want to use a linear frequency filterbank with rectangular channels as a simple starting point. This can be extended to a mel-scaled filterbank and to then incorporate triangular shaped channels to ultimately produce mel-frequency cepstral coefficients (MFCCs). You may also consider augmenting the feature vector with its temporal derivatives as this may increase recognition accuracy. The different configurations should provide you with interesting designs that you can evaluate.

The feature extraction code should take as input a speech file (for example `speech001.wav`) and output a file containing, for example, MFCC vectors (for example `speech001.mfc`).

## 3. Acoustic modelling

Acoustic modelling is where the acoustic properties (as represented by the feature vectors) are modelled.

For this assignment, hidden Markov models (HMMs) will be used as the acoustic model. You do not need to implement code for the HMMs. Instead the HMM toolkit (HTK) will be used. The scripts from Labs 3 allow you to input your feature vectors and label files into HTK from which acoustic models will be created. For each item in your vocabulary (and silence) an acoustic model should be trained. The training will use the HInit and HRest tools of HTK. You will also be able to change the topology of the HMMs and evaluate how this changes the recognition accuracy.

### 4. Noise compensation

Noise can be added to the clean speech samples to create noisy speech which is more representative of real-world use of speech recognition systems. Adding noise to the speech will reduce the decoding accuracy and increase confusions. To mitigate this, some form of noise compensation may be needed. Different methods can be tested such as applying spectral subtraction to the feature extraction process or training the speech models on noisy speech (matched models). The effect of different types of noise can be investigated and different signal-to-noise ratios (SNRs).

### 5. Testing and evaluation

Once you have a set of acoustic models, training of the speech recogniser is complete and it can now be tested. Testing involves passing a new speech file (in the same feature format as the training data) to the speech recogniser and letting it recognise the speech. This uses the HVite tool of HTK. This will recognise one speech utterance, but for a proper evaluation many speech files (at least 10 of each word in the vocabulary) should be tested.

Therefore, a new set of speech files should be collected (for example a further 10 or 20 examples of each word in the vocabulary) and input into the speech recogniser. The recogniser will output a `.rec` file which will contain the word that it thinks was spoken. By comparing the `.rec` file with the true label for the utterance (i.e. the label file - `.lab` file) you can determine whether the recogniser was correct or not. For example, comparing the true, or reference, label stored in `speech001.lab` with the speech recogniser output `speech001.rec`. The HTK tool HResults will carry out this analysis for you and output both the overall accuracy of the speech recogniser as a percentage and a confusion matrix that shows which word confusions took place.

Within the evaluation you can examine the effect of different configurations of feature extraction and HMM topology. This may include different numbers of filterbank channels, different spacing of the channels, the frame rate, etc. You may also want to test your speech recogniser in noisy conditions (for example, factory noise, babble noise, etc) and under different signal-to-noise ratios (SNRs) to examine how the noise affects recognition accuracy. You may then want to examine how noise compensation is able to effect performance in noisy conditions. For all tests, include text to explain what is happening and why you think this is the case.

## Relationship to formative assessment

Formative assessment takes place during all lab classes through discussion of your analysis, designs and implementations. These labs underpin the coursework and relate directly to the different parts.

## Deliverables

The assessment covers two parts and represents two of the assessed components of CMP-6026A:

*Technical report (CW001)*

You should produce a technical report that describes the data collection, design, implementation and evaluation of your speech recogniser. The first part of the report should be written jointly by both people in the group and should contain details on the speech collection and annotation, feature extraction and acoustic modelling. The second part of the report should be written individually and describe the evaluation of the speech recogniser in terms of its performance under different test conditions. The report should be no longer than 8 pages of A4. You should submit individual reports, which will have the same first part but individual second (evaluation) parts. Submission is electronically through Blackboard.

*Practical demonstration of the recogniser (CW002)*

The practical demonstration will take place in the Electronics Lab. In the practical demonstration you will be asked to say a sequence of names that you will then decode using your speech recogniser. You will also be expected to discuss your system and justify design decisions. Each group will have up to 10 minutes for the demonstration and individual marks will be awarded, so ensure both people in the pair make a roughly equal contribution.

## Resources

You will need to use audio/visual recording equipment/software, MATLAB, HTK, SFS as used in the lab classes. These resources have been introduced in the lectures and lab classes.

There will be a briefing session for this coursework in Week 4.

## Marking scheme

Marks will be allocated as follows for the two assessed components:

Technical report (75%)

- Speech collection and annotation methodology (5%)
- Design and justification of feature extraction (20%)
- Design and justification of acoustic modelling (5%)
- Noise compensation (5%)
- Evaluation of the performance of the speech recogniser (30%)*
- Structure and technical writing style of report (10%)*

Demonstration and discussion (25%)

- Organisation of demonstration and discussion/question answering (25%)*

* The evaluation/structure parts of the technical report and the demonstration/discussion will be awarded as individual marks while the remainder will be a joint mark.

**Note** - it is expected that both people in the pairing will make approximately equal contributions. If it becomes apparent that this is not the case, then for fairness, the distribution of marks allocated may be adjusted.

**UNIVERSITY OF EAST ANGLIA**
**School of Computing Sciences**

**MODULE: CMP-6026A**

**ASSIGNMENT TITLE: Design, implementation and evaluation of a speech recognition system**

| | |
|---|---|
| **Names:** | **Registration Nos:** |

| Section | Part and weight | Mark | Comments |
|---|---|---|---|
| Report (75%) | Speech collection and annotation methodology (5%) | | |
| | Design of justification of feature extraction (20%) | | |
| | Design and justification of acoustic modelling (5%) | | |
| | Noise compensation (5%) | | |
| | Evaluation of speech recognition performance (30%) | | |
| | Structure and technical writing style of report (10%) | | |
| Demo (25%) | Organisation of demonstration and discussion/question answering (25%) | | |

**Provisional Mark Awarded:** **Date:**

All marks are provisional until confirmed by the Board of Examiners.