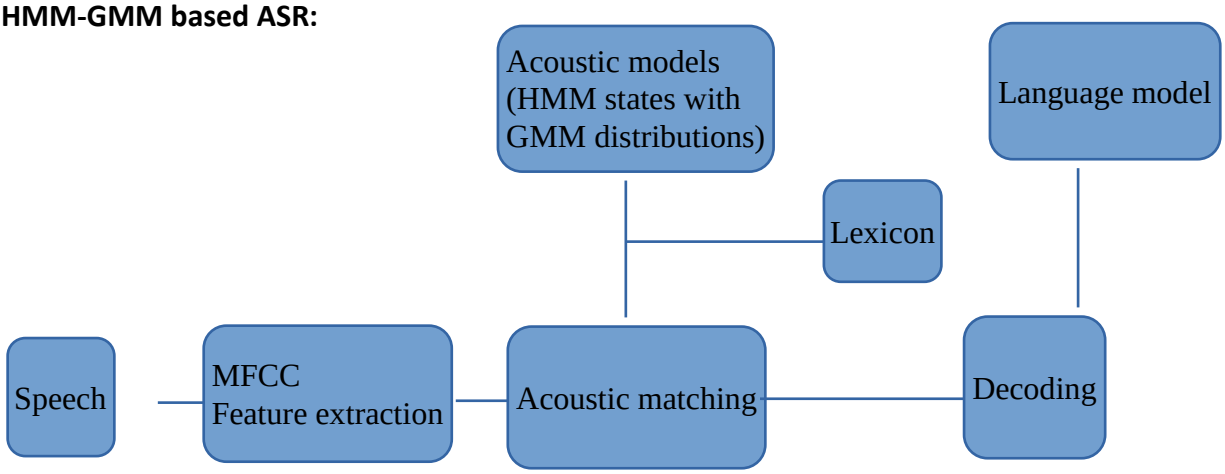
Automatic Speech Recognition – report
194102311 – Mathav raj

Speech recognition is a pattern recognition task involving four modules working together to decode a speech signal into the most likely text output. Here we use Kaldi toolkit to perform the training and testing of a speech dataset.

HMM-GMM based ASR:



- MFCC - speaker independent feature vectors for exclusively recognising the audio
- Acoustic model - probability of acoustic observations given a word sequence (HMM - GMM)
- Lexicon - dictionary of valid phoneme sequences in language (manually provided)
- Language model – Prior probabilities for word sequence (trained by an external text corpus)

LM experiments:

As a first step we trained and visualised the perplexity scores of a N-gram language model.

Ngram	Log probability	Perplexity with sentence markers	Perplexity without sentence markers
2	-17312.52	424.1194	524.7542
3	-17244.6	414.1713	512.0178
5	-17244.24	414.1196	511.9516

Perplexity of an N gram model measures how 'surprising' the test data looks to the model developed. Lower the perplexity score better the model is.

ASR Experiment:

- Three kinds of models are trained and verified with training and development data and tested with test data
- Test data folder - G9
- Language model selected - Bigram
- While training the acoustic model for every epoch we go through alignment steps to align the acoustic model so far created and the reference transcript. Alignment is needed to tell the model that a particular phoneme might occupy multiple consecutive frames.

HMM - GMM acoustic model:

HMM tracks the sequential information of the phonemes. Under each HMM state the observation probabilities of a phoneme is modelled by a GMM. GMM is flexible in modelling the general non unimodal continuous distribution of MFCC vectors due to various factors like phonetic context.

Depending on the state modelled by HMM we have,

- Monophone acoustic model - phoneme level HMM, no contextual information about the preceding or following phone
- Triphone acoustic model – contextual, built by tying up monophone models by a decision tree

DNN acoustic model:

A neural network with backpropagation algorithm can learn the subtle variations in phonemes as good as GMM. Here, the senones (context dependent units) of the triphone HMM states are modelled by the output layer of a DNN.

Performance on development data:

Model	Word error rate
Monophone	68.33
Triphone	47.43
DNN	45.26