MP2 Walkthrough
HMM Speech Recognition

ECE 417 – Multimedia Signal Processing
Fall 2018

Goal

- Implement a speech recognizer using Hidden Markov Model(HMM) to recognize certain words

Data Corpus

- 100 different audio files:
  - 4 speakers: mh, ls, dg, yx
  - 5 words: "CNN", "DNN", "ASR", "TTS" and "HMM"
  - 5 utterances of each word per speaker

Overview

- Extracting audio features
- Splitting training and testing data
- Training Gaussian HMM model for speech recognizer
- Evaluating your HMM model

Extracting audio features

- Extract the features to represent the audio recordings
- You are provided with the MFCC features for each audio recording
- BONUS POINTS! Up to 10%
- New feature set other than MFCC
  - Implement, report the results, and beat reference implementation accuracy results

Splitting training and testing data

- Speaker dependent experiment
  - Training: first 4 utterances of each word, from each of the 4 speakers (4x4x5=80)
  - Testing: fifth utterance from each speaker (4x5=20)
- Speaker independent experiment
  - Training: all utterances from speakers dg, ls, and yx (3x5x5=75)
  - Testing1: all utterances from speaker mh (5x5=25)
  - Testing2: all utterances from you (5x5=25)

Training the Gaussian HMM

- Recap of HMM:
  - A HMM is a statistical model for a time-varying process
  - The entire model represents a probability distribution over the sequence of observations
    - It has a specific probability of generating any particular sequence
  - It consists of two components
    - A Markov chain that specifies how many states there are, and how they can transition from one state to another
    - A set of probability distributions, one for each state, which specifies the distribution of observation in that state
  - HMM Parameters
    - π - initial state distribution
    - A - state transition matrix
      - Aij is the probability that when in state i, the process will move to j
    - B - observation matrix
      - Probability of data produced from any state
      - In this lab, model the observation matrix as Gaussian (μ,σ)

$$A = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0 & 0.7 & 0.3 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

Training the Gaussian HMM

- Learn the HMM parameters (π, A, μ, σ) from observation sequences/training utterances
- Approach: forward-backward/EM algorithm to optimize the parameters
- Initialization:
    - π – uniform distribution across 5 states
    - A - [0.8 0.2 0 0 0; 0 0.8 0.2 0 0; 0 0 0.8 0.2 0; 0 0 0 0.8 0.2; 0 0 0 0 1]
    - μ - mean across the audio features for that word
    - σ - co-variance matrix across the audio features for that word

Training the Gaussian HMM

- BONUS POINTS! Up to 10%
- The observation matrix can also be modeled as likelihood function other than Gaussian
- Examples: GMM, KNN, NN…
- Write your own code to integrate the function with the HMM, and beat baseline
- Partial credit(8%) possible with explanation

Evaluating the model

- Given your trained model parameters of each word, compute the likelihood of word utterance in the test set
- Classify the utterance as the word with maximum likelihood
- Report the average classification accuracy on all the word utterances in your testing data

Results

- Confusion matrix: a 5x5 matrix in which the (m,n)th element specifies the conditional probability that the recognizer chose the nth word, given that the mth word was correct.
- Overall recognition accuracy for each of the three experiments

Results

- Confusion matrix: a 5x5 matrix in which the (m,n)th element specifies the conditional probability that the recognizer chose the nth word, given that the mth word was correct.
- Overall recognition accuracy for each of the three experiments

Predicted word

Ground truth word

	ASR	CNN	DNN	HMM	TTS
ASR	1	0	0	0	0
CNN	0	1	0	0	0
DNN	0	0.25	0.75	0	0
HMM	0	0	0	1	0
TTS	0	0	0	0	1

Results

- Confusion matrix: a 5x5 matrix in which the (m,n)th element specifies the conditional probability that the recognizer chose the nth word, given that the mth word was correct.
- Overall recognition accuracy for each of the three experiments

Predicted word

Ground truth word

	ASR	CNN	DNN	HMM	TTS
ASR	1	0	0	0	0
CNN	0	1	0	0	0
DNN	0	0.25	0.75	0	0
HMM	0	0	0	1	0
TTS	0	0	0	0	1

Results

- Confusion matrix: a 5x5 matrix in which the (m,n)th element specifies the conditional probability that the recognizer chose the nth word, given that the mth word was correct.
- Overall recognition accuracy for each of the three experiments

Predicted word

Ground truth word

	ASR	CNN	DNN	HMM	TTS
ASR	1	0	0	0	0
CNN	0	1	0	0	0
DNN	0	0.25	0.75	0	0
HMM	0	0	0	1	0
TTS	0	0	0	0	1

Turn In

- Report
  - Include the confusion matrix and overall recognition accuracy for each of three experiments
  - Include your analysis of comparisons between the outputs
  - [Optional] Include your results for extra credit in the end
  - File names must be <Lastname>_<Firstname>_report.pdf
- Code
  - Readme file
  - File names must be <Lastname>_<Firstname>_code.zip
  - Do not upload the data corpus
- Submission
  - Submit your report (PDF) and codes (zip) to Compass
  - Teams will submit a single report but make sure that all names are included in the report