Speech recognition using HMM

Assignment summary

In recognizing speech the role of HMM is clearly visible. We prepare HMM Model for each speech signal and train them. When the models are trained we can use them to evaluate a new speech signal using evaluation method.

Thus by comparing the likelihood value computed on each model for the speech signal we can predict the speech signal belonging to that class which likelihood measure is highest i.e. difference is minimum.

Following steps are to be taken -

- Record 5 types of speech (words in .wav format) signals having time duration 03 sec. Each having five samples. 3 for training and 2 for testing.
- Every speech signal is recorded at 128 kbps bit rate and 8000 Hz frequency (you may use audacity software: http://audacity.sourceforge.net/) Or Windows sound recorder.
- Convert each signal of same size (means number of sample value in each speech signal should be) same) (nX1) (down sampling, resize).
- Arrange all samples row wise i.e 1Xn. Total number of rows will be (15Xn). Each will be represented as a vector.
- **Generating codebook (This is most important)**

1. **Find out the optimized cluster heads (say C1,C2,……Ck)**
2. **Find out the belonging of each class sample to these centroids and name it accordingly.**

- Creating a HMM model for that speech signal
- Training the HMM model using symbol sequence
- Testing the model

Stepwise implementation for Training

Step1:

- Record 5 types of speech words. Each word having five samples. 3 for training and 2 for testing. Convert each signal of same size means number of sample value in each speech signal is same (nX1) where n =total number of amplitudes for each word of duration 2.5-3.5 sec.
- Arrange all samples row wise i.e 1Xn. Total number of rows will be (15Xn). Each will be represented as a vector S.

S = $s_{11}s_{12}s_{13}\ldots\ldots\ldots\ldots s_{1n}$ = s1 (First sample vector)

$s_{21}s_{22}s_{23}\ldots\ldots\ldots\ldots s_{2n}$ = s2

……

……

$s_{151}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots s_{15n}$ = s15 (last , i.e, 15th sample vector)

Each represented as a vector.

Steps of implementation contd…

Step2: Code book generation

- LBG algorithm is used to generate codebook which iteratively uses K-means to generate sets and their centroid. It calculate the distortion in the data with generated codebook and since it iteratively operates K-means it selects the one with minimum distortion.
- Choose k=5 clusters with 5 centroids.
- Randomly select 5 vectors as a centroid of each cluster represented as c_i. Which is a initial codebook c0.
- Find the Euclidean distance of each vector to each of the centroids.

$$D(s, c) = \sqrt{\sum_{i=1}^{n} (s_i - c_i)^2}$$

- Search the nearest vector which has the minimum distance D.
- New centroids will be calculated by taking average of vectors belongs to a particular cluster.
- This process is continued until the difference between two centroids is less then 0.001 or 5 to 10 iterations .
- Final codebook is represented as cf.

Step 3: Generation of HMM model for every speech signal.

- Consider the first speech signal with 3 samples than arrange them row wise similar to step 1 denoted as t where t_{3Xn} is the sample vector of first speech signal.
- Calculate the Euclidean distance of each vector to the final generated codebook.

$$D(t, c) = \sqrt{\sum_{j=1}^{3} \sum_{i=1}^{n} \left(t_{(j,i)} - cf_{(1,i)}\right)^2}$$

- Replace the vector with its nearest distance codebook which implies $t_{(j,i)}$ is replaced by $cf_{(1,i)}$. Do it for all the vectors.
- In this manner the observation sequence is generated. Here we consider the nearest integer for representing the observation sequence.
- Now Baum welch algorithm is applied. Assume number of states is 5 : s1 s2 s3 s4 s5.
- Randomly take the transition matrix a_{ij} where $\sum a_{ij} = 1$
- Randomly consider the emission matrix b_{ij} with 3 observation sequences where $\sum b_{jk} = 1$, here number of observation is equal to the number of cluster present in the LBG Algorithm.
- Randomly select Initial probability of all five states p(s1),p(s2), p(s3) ,p(s4),p(s5)= 1/5.

- Find the state sequence for a particular observation sequence obtained using codebook p(s/o).
- Update a_{ij}, b_{jk} and initial probability π_i using Baum welch algorithm.
- When the log likelihood probability difference in two consecutive iteration is less than some threshold value the learning stops(<0.001).
- This is HMM model for first speech signal H1(A, B, π).
- In similar way HMM model for each speech signal is generated.
- Store these models for further use in testing phase.

Implementation for testing

Testing Steps:

- From the training we have 5 different models for each word with values A(State Transition Matrix), B(Emission Matrix), π (Initial Probabilities).
- Take the test sample then find out the observation sequence for that corresponding word (with the help of codebook) let the sample belongs to O_{test} where test$\in$[1 5].
- After that find max{ P(O_{test}/ $\boldsymbol{\lambda}$) } where $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \boldsymbol{\lambda}_4, \boldsymbol{\lambda}_5$

  using viterbi algorithm. For example If h1 has the maximum value, sample will belongs to h1 (first class: hello).
- $\boldsymbol{\lambda}_1$ (A, B, π) is the first HMM model for first class (Hello).
- $\boldsymbol{\lambda}_2$ (A, B, π) is the second HMM model for second class (How).
- In similar way $\boldsymbol{\lambda}_5$ (A, B, π) is the fifth HMM model for fifth class.