AI1103 Project

Ayush Kumar Singh- AI20BTECH11028

Comparision of Various Techniques for Speaker Recognition

Abstract

In this presentation, a comparision on different speaker recognition technique is shown. The techniques are vector quantization (VQ) using Linde Bozo Gray(LBG), gaussian Mixture Model (GMM) using EM algorithm and Hidden Markov Model(HMM). VQ adds the method of considering a large group of feature vectors of a known user and generating a smaller group of feature vectors that signifies the centroid of spreading, i.e. points set apart, so as to reduce the distance between the points. The GMM can be represented in the form of a summation of the VQ model where the clusters are overlying. HMM is a finite group of states, each of which is united with a probability distribution.

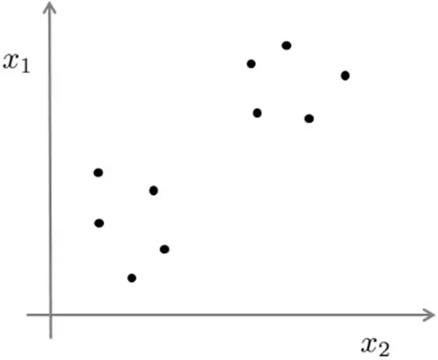
A brief Overview on speaker recognition

All speakers recognition systems contain three main modules:
(a) Acoustic Processing (b) Feature Extraction (c) Feature Matching.

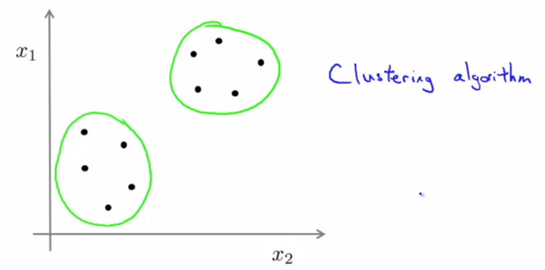
1. In acoustic processing, an analog signal is received from a speaker and regenerates it into a digital signal for additional digital processing.
2. The signal is then fed into the spectral instrument for feature extraction.
3. Feature matching includes the particular method to spot the unknown speaker by analysis the extracted features from voice input with those from a group of well-known speakers.

Some important concepts

Clustering: grouping of data



Let $X = [X_1, X_2]$ be a feature vector. Above figure shows distribution for a sample of data.



Now, upon using a clustering algorithm we can get a group of clusters for these non labelled data.
This is useful since in methods like VQ and GMM we extract the features of a sample of data and try to make clusters based upon various algorithm which is explained in detail later.

Vector Quantization (VQ)

In VQ technique, a vector of a large space is mapped to a fixed number of regions in that space. These types of regions are called clusters and these are represented by its center that is also known as centroid.

Vector quantization as a Data Compressor

For the process of speech coding the step of quantization is mandatory for reducing the bits number that have been used to characterize the samples of a signal.

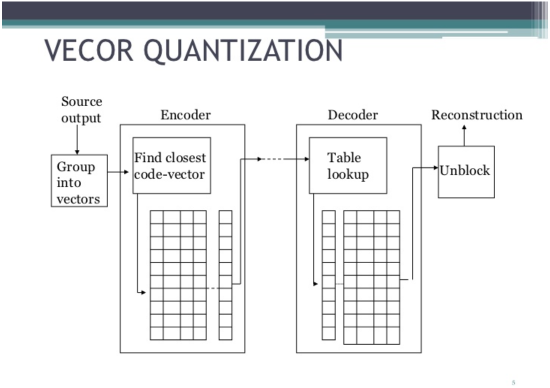
In vector quantization we prepare a code book and for storing a vector we find the minimum distortion of the given vectors with code vectors to find the index of the vector.

Distortion of a vector is given by:

$$Distortion = |X - Y_i|^2 \tag{1}$$

where Y_i denotes the code vector and X is the given feature vector

Working of VQ



VECOR QUANTIZATION

1. We first find the minimum distortion for a given vector from codebook and assign an index to the given vector.
2. To retrieve the value of a vector we simply use to lookup table to find the vector value from the obtained index.

Formation of Codebbok

LBG algorithmic rule is employed for clump a group of L training vectors into a group of M codebook vectors is being employed. The LBG algorithm is used to design the codebook which is further used in clustering.

1. Assume a random vector as centroid for first cluster
2. Split the centroid which doubles the size of present codebook Y_n. It can be represented as per rule:

   $$Y_n^+ = Y_n(1+\beta), Y_n^- = Y_n(1-\beta) \tag{2}$$

   where the range of n is from 1 to the size of the present codebook and β is known as splitting parameter.
3. Assign each training vector to a cluster related with the nearest code vector through nearest neighbor search procedure.
4. Update each centroid of a codebook.
5. At last the distortion of each training vector is measured by repeating third and fourth step until the distance lies below threshold value.
6. Repeat above steps until codebook of size M is obtained.

Example

QUESTION

For a given vector [2,3] and cluster centroids [0,0], [1,0] and [2,0], find the cluster in which the given vector belongs.

SOLUTION

To find the cluster we have to calculate the distortion of the given given vector with each cluster centroid

$$d_1 = (2-0)^2 + (3-0)^2 = 13 \tag{3}$$

$$d_2 = (2-1)^2 + (3-0)^2 = 10 \tag{4}$$

$$d_3 = (2-2)^2 + (3-0)^2 = 9 \tag{5}$$

Since d_3 is minimum, the given vector belongs in the third cluster.

Gaussian Mixture Model

Mixture Model

A mixture model is a probabilistic model in which probability distribution is represented as a combination of simpler component distribution.
A gaussian mixture model on the other end is a mixture model in which each component is assumed to be gaussian.
Gaussian Mixture Model is also a clustering algorithm in which feature vector of training speech data is organised in form of cluster and speech recognition is done on the basis of cluster in which the given feature vector belong.

Gaussian Mixtures

1. Linear super position of Gaussians

$$p(x) = \sum_{K=1}^{K} \Pi_k N\left(x|\mu_k, \sum_k\right) \tag{6}$$

   where k= number of gaussians
   π_k=weight of each gaussian,
   μ_k=mean,
   $\sum_k$=covariance.

2. Normalization and positivity:

$$0 \leqq \mu_k \leqq 1, \sum_{k=1}^{n} \pi_k = 1 \tag{7}$$

3. log - likelihood:

$$lnp(x_n) = \sum_{n=1}^{N} ln\left(\sum_{K=1}^{K} \Pi_k N\left(x|\mu_k, \sum_k\right)\right) \tag{8}$$

   where N= number of data points

Why a new clustering algorithm?

Limitation of LBG algorithm

1. An instance of data belong to only one cluster
2. overlapping clusters not possible
3. Not a probabilistic model.
4. Prefers equal sized cluster

Simulation

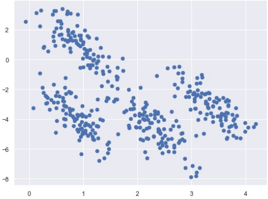


Figure: data points

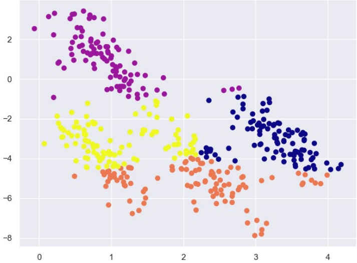


Figure: clustering with k means(LBG)

cntd.

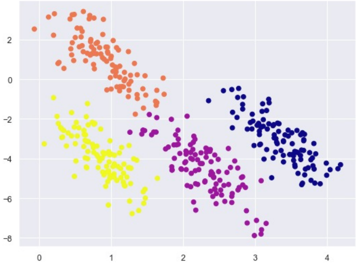


Figure: clustering with Gaussian Mixture Model

Hence, we can easily see that due to equal size limitation of LBG algorithm the clusters are not suitable for a complex arrangement of data points and gaussian mixture model is preferable.

Latent variabe: posterior probability

1. We can think of mixing coefficient as prior probabilities for each component.
2. For a given value of x we can evaluate the corresponding posterior probabilities, called responsibilities

$$\gamma(x) = P(K|x) = \frac{p(k)p(x|k)}{p(x)} \tag{9}$$

$$= \frac{\Pi_k N\left(x|\mu_k, \sum_k\right)}{\sum_{j=1}^{K} \Pi_j N\left(x|\mu_j, \sum_j\right)} \tag{10}$$

where
$\pi_k = \frac{N_k}{N} and N_k is the effective number of points assigned to a given cluster.$

Expectation Maximization

1. EM algorithm is an iterative optimization technique which is operated locally.
2. Estimation step : For given parameter value we compute the expected value of latent variable(responsibility).
3. Maximization Step: Update the parameters of the model based on the value of the latent variable calculated.
4. Given a Gaussian Mixture model our goal is to maximize the likelihood function with respect to the parameters comprising of mean and covariance of the component and mixing coefficients(weight).

EM algorithm

1. initialize the mean, covariance and mixing coefficient of the gaussians and evaluate the value of log likelihood.
2. Evaluate the responsibility from the given parameter(Estimation step)

$$\gamma(x) = P(K|x) = \frac{p(k)p(x|k)}{p(x)} \tag{11}$$

$$= \frac{\Pi_k N\left(x|\mu_k, \sum_k\right)}{\sum_{j=1}^{K} \Pi_j N\left(x|\mu_j, \sum_j\right)} \tag{12}$$

3. Re evaluate the value of parameters using the calculated responsibility

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)x_n}{\sum_{n=1}^{N} \gamma_j x_n} \tag{13}$$

cntd.

$$\sum_{j} = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu j)(x_n - \mu j)^T}{\sum_{n=1}^{N} \gamma_j x_n} \tag{14}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n) \tag{15}$$

Now evaluate Log likelihood, if there is no convergence return to step 2 and continue the iteration.

Simulation

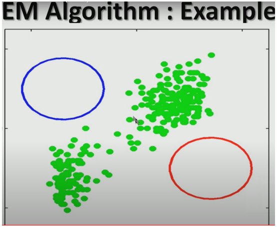


Figure: clustering with Gaussian Mixture Model

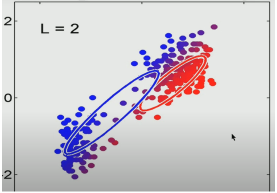


Figure: clustering with Gaussian Mixture Model

cntd.

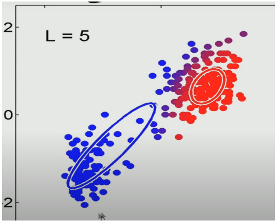


Figure: clustering with Gaussian Mixture Model

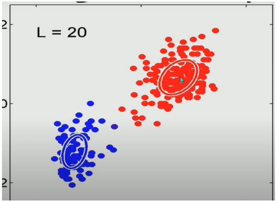


Figure: clustering with Gaussian Mixture Model

Hidden Markov Model

This model basically refers to a model where system that is modelled by using Markov process in which the states always be unobserved. It is characterized as the easiest dynamic Bayesian process.
Markov model refers to model in which observer can view the state directly, and thus the state transition probabilities are the only featured parameters, the HMM, differs in the fact that the observer cannot directly view the state, but the output is visible in the form of token or data and this output which is depends on the state.

HMM for pattern matching

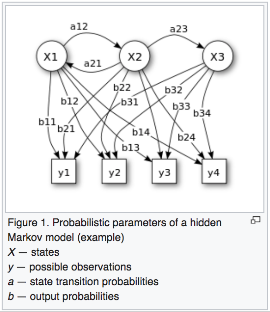


Figure 1. Probabilistic parameters of a hidden Markov model (example)

X — states

y — possible observations

a — state transition probabilities

b — output probabilities

Figure: An example of hidden markov model

For the above hidden markov model, suppose we have to find the sequence of state which is most probable for the given observed state $(y_1, y_2 ... y_n)$

cntd.

For that we have to find the sequence of state with maximum probability given Y:

$$max_{X=x1,x2...xn}P(X = X_1, X_2..., X_n|Y = Y_1, Y_2, ...Y_n) \tag{16}$$

$$= \frac{P(Y|X)P(Y)}{P(X)}\text{From bayes theorem} \tag{17}$$

$$Also, P(Y|X) = \prod P(Y_i)P(X_i) \tag{18}$$

$$P(X_i) = \prod P(X_i|X_{i-1}) \tag{19}$$

$$\text{Therefore we get , } max_{X=x1,x2..xn} \prod P(Y_i|X_i)P(X_i|X_{i-1}) \tag{20}$$

HMM Block Diagram

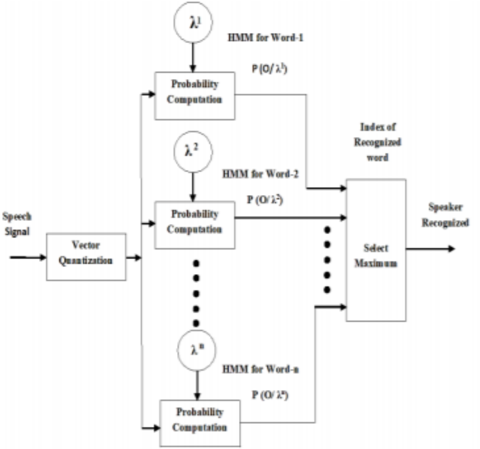


Figure: Block diagram demonstrating HMM

Advantages Of HMM

1. HMM covers both temporal and spectral changeability in flexible manner.
2. Not only an allophone, phoneme, syllable, or a word can be characterised by a HMM, but an entire sentence can be represented by 1 large composite HMM.
3. HMM formulation can be applied not only to English but equally well to the other languages.

conclusion

1. The performance of VQ using LBG to measure and create the codebook, HMM with suitable algorithm to find the all possible probability of given condition and GMM with EM algorithm to find the finest value of responsibities (mean, standard deviation and the weighting factor of curve) are compared for speaker recognition.
2. It is confirmed that the feature model using GMM is superior to VQ and HMM.
3. The average recognition rate attained for GMM is 99.22, for HMM is 97.26 and for VQ with LBG is 91.43. These given data are taken based on theoretical knowledge of speaker recognition from various proposed papers.