EEE 6110 Speech Processing.

Dr. Ciira Maina
ciira.maina@dkut.ac.ke

10th April, 2019

Speech Recognition

- The goal is to convert the speech signal into text
- Applications include improved human-machine interaction

Speech Recognition

- The steps involved in building a speech recognition system include
  - Select a feature set
  - Choose the recognition vocabulary, basic speech sounds
  - Train the acoustic and language models
  - Evaluate performance

See Figure 9.2 Rabiner and Schafer

Feature Extraction

See Figure 9.3 Rabiner and Schafer

Acoustic and Language Modelling

- Acoustic modelling requires accurately labelled sequences of speech utterances
- The recordings are segmented according to the transcription
- Language modelling requires text strings reflecting the syntax of the language

Performance Measures

- Accuracy
- Word error rate
- Sentence error rate

Mathematical Description of ASR

- Mathematically we have

$$\hat{W} = \arg\max_{W} P_A(X|W) P_L(W) \tag{1}$$

- X is a sequence of acoustic observations

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \tag{2}$$

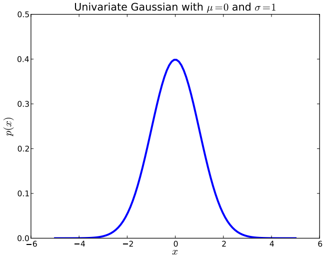
- W is a sequence of words

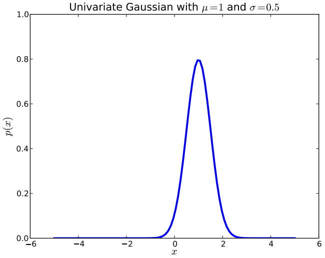
$$W = w_1, \ldots, w_M \tag{3}$$

Acoustic Modelling

- The Gaussian distribution function for a 1D variable is given by
$$p(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$
- The distribution is governed by two parameters
  - The mean μ
  - The variance σ^2
- The mean determines where the distribution is centered and the variance determines the spread of the distribution around this mean.

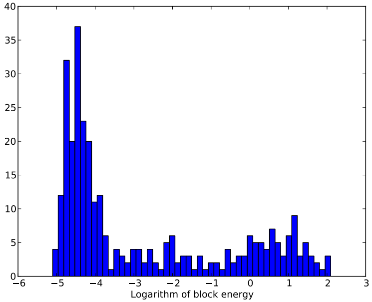
Acoustic Modelling





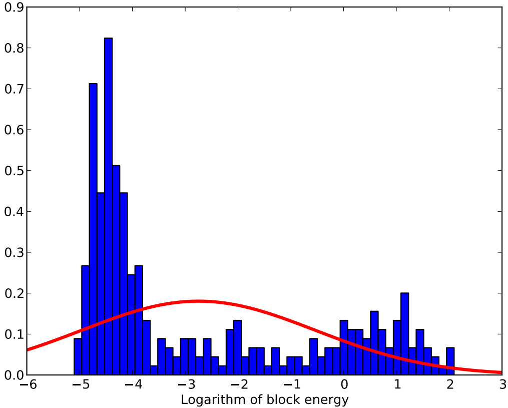
Acoustic Modelling - VAD Example

- Voice activity detection is a useful signal processing application
- It involves deciding whether a speech segment is speech or silence
- We divide the speech into short segments and compute the logarithm of the energy of each segment.
- We see that the log energy shows distinct clusters.



Acoustic Modelling - VAD Example

- A single Gaussian does not fit the data well



Gaussian Mixture Models, Theory

- The Gaussian density can not be used to model data with more than one distinct 'clump' like the log energy of the speech frames.
- Linear combinations of more than one Gaussian can capture this structure.
- These distributions are known as Gaussian Mixture Models (GMMs) or Mixture of Gaussians

Gaussian Mixture Models, Theory

- The GMM density takes the form

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k)$$

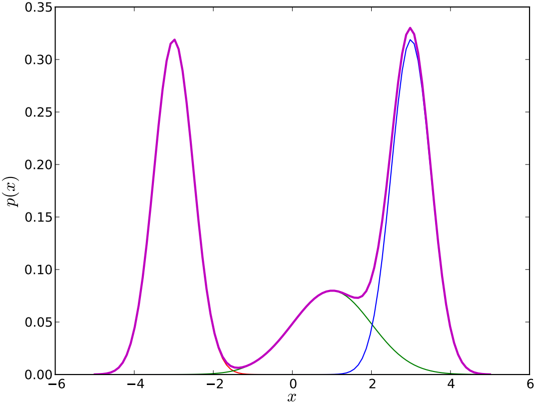
- π_k is known as a mixing coefficient. We have

$$\sum_{k=1}^{K} \pi_k = 1$$

  and $0 \leq \pi_k \leq 1$

Gaussian Mixture Models, Theory

- A GMM with three mixture components



Gaussian Mixture Models, Theory

- The mixing coefficients can be viewed as the prior probability of the components of the mixture
- We can then use the sum and product rules and write

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k)$$

- Where

$$p(k) = \pi_k$$

and

$$p(x|k) = \mathcal{N}(x|\mu_k, \sigma_k)$$

Gaussian Mixture Models, Theory

- Given an observation x, we will be interested to compute the posterior probability of each component that is $p(k|x)$
- We use Bayes' rule

$$
\begin{aligned}
p(k|x) &= \frac{p(x|k)p(k)}{p(x)} \\
&= \frac{p(x|k)p(k)}{\sum_i p(x|i)p(i)}
\end{aligned}
$$

- We can use this posterior to build a classifier

Gaussian Mixture Models, Learning the model

- Given a set of observations $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ where the observations are assumed to be drawn independently from a GMM, the log likelihood function is given by

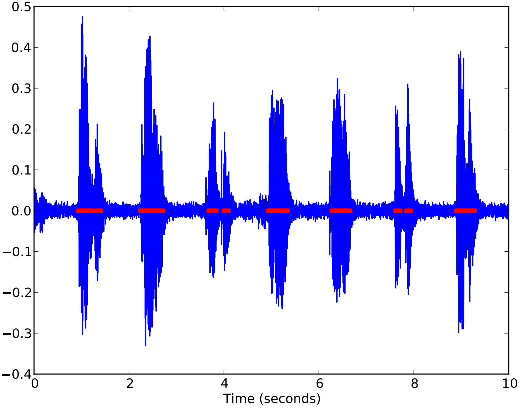
$$\ell(\theta; \mathbf{X}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k) \right\}$$

  where $\theta = \{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2\}$ are the parameters of the GMM.
- To obtain a maximum likelihood estimate of the parameters, we use the expectation maximization (EM) algorithm

Gaussian Mixture Models, Returning to the VAD Example

- In the VAD example we use the implementation of EM in scikit-learn.
- We can then compute the posterior probability of all segments belonging to the component with the highest mean.
- Segments where this probability is greater than a threshold can be classified as speech.

Gaussian Mixture Models, Returning to the VAD Example



Sequential Models

- Often the i.i.d assumption is poor
- In a speech signal, frames are not i.i.d
- One can predict current values based on past values
- In Markov models, future predictions are independent of all except the most recent observations

First Order Markov Models

- In general

$$p(X) = p(\mathbf{x}_1) \prod_{i=2}^{T} p(\mathbf{x}_i | \mathbf{x}_{i-1}, \dots, \mathbf{x}_1) \tag{4}$$

- If we assume that

$$p(\mathbf{x}_i | \mathbf{x}_{i-1}, \dots, \mathbf{x}_1) = p(\mathbf{x}_i | \mathbf{x}_{i-1}) \tag{5}$$

- We obtain the first order Markov chain

$$p(X) = p(\mathbf{x}_1) \prod_{i=2}^{T} p(\mathbf{x}_i | \mathbf{x}_{i-1}) \tag{6}$$

Hidden Markov Models

- To allow for richer structure, we introduce latent (hidden) variables $\{\mathbf{z}_n\}$
- The model is now given by the joint distribution

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{z}_1, \ldots, \mathbf{z}_T) \tag{7}$$

- The latent variables are assumed to form a first order Markov chain and the joint distribution becomes

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{z}_1, \ldots, \mathbf{z}_T) = p(\mathbf{z}_1) \prod_{i=2}^{T} p(\mathbf{z}_i|\mathbf{z}_{i-1}) \prod_{i=1}^{T} p(\mathbf{x}_i|\mathbf{z}_i) \tag{8}$$

- When the latent variables are discrete, we obtain a hidden Markov model

Hidden Markov Models

- At a single time slice, the model corresponds to a mixture distribution with component distributions given by $p(\mathbf{x}|\mathbf{z})$
- The latent variables $\mathbf{z}_i$ are multinomial variables
- We adopt a 1-of-K encoding scheme
- The transition probabilities are represented in a transition matrix $\mathbf{A}$
- $a_{jk} = p(z_{ik} = 1 | z_{i-1,j} = 1)$ where $0 \leq a_{jk} \leq 1$ and $\sum_k a_{jk} = 1$
- We have

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_T) = p(\mathbf{z}_1) \prod_{i=2}^{T} p(\mathbf{z}_i | \mathbf{z}_{i-1}) \tag{9}$$

Hidden Markov Models

- We have
$$p(\mathbf{z}_i|\mathbf{z}_{i-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} a_{jk}^{z_{i-1,j} z_{ik}}$$
- and
$$p(\mathbf{z}_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$$
- The distributions of the observed variables depend on parameters ϕ. That is $p(\mathbf{x}_i|\mathbf{z}_i, \phi)$

Hidden Markov Models

- The joint distribution of latent variables and observed variables is given by

$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{z}_1|\pi)\left(\prod_{i=2}^{T} p(\mathbf{z}_i|\mathbf{z}_{i-1}, \mathbf{A})\right)\prod_{i=1}^{T} p(\mathbf{x}_i|\mathbf{z}_i, \phi)$$

- Where $\theta = \{\pi, \mathbf{A}, \phi\}$
- Give data, the parameters θ are estimated using maximum likelihood

Readings

- HAH - Chapter 8
- RS - Chapter 9