Speaker Recognition

David Cinciruk

2/24/2012

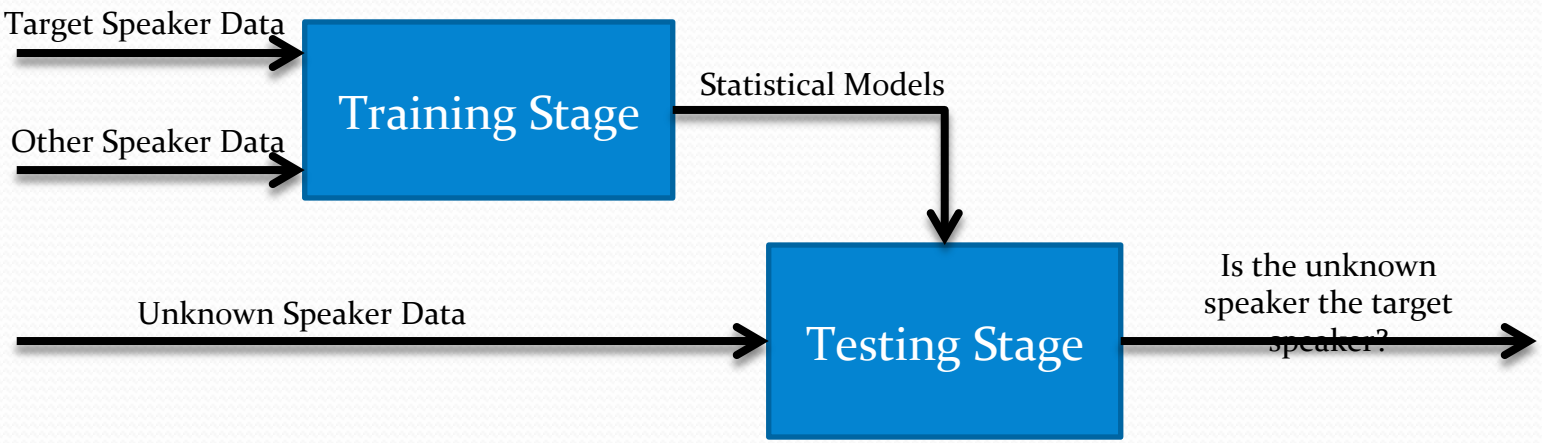
ASPITRG Group Meeting

Table of Contents

- The Basics of Speaker Recognition
- Creating MFCCs
- Training the UBM
- Adapting the UBM
- Scoring
- Conclusion

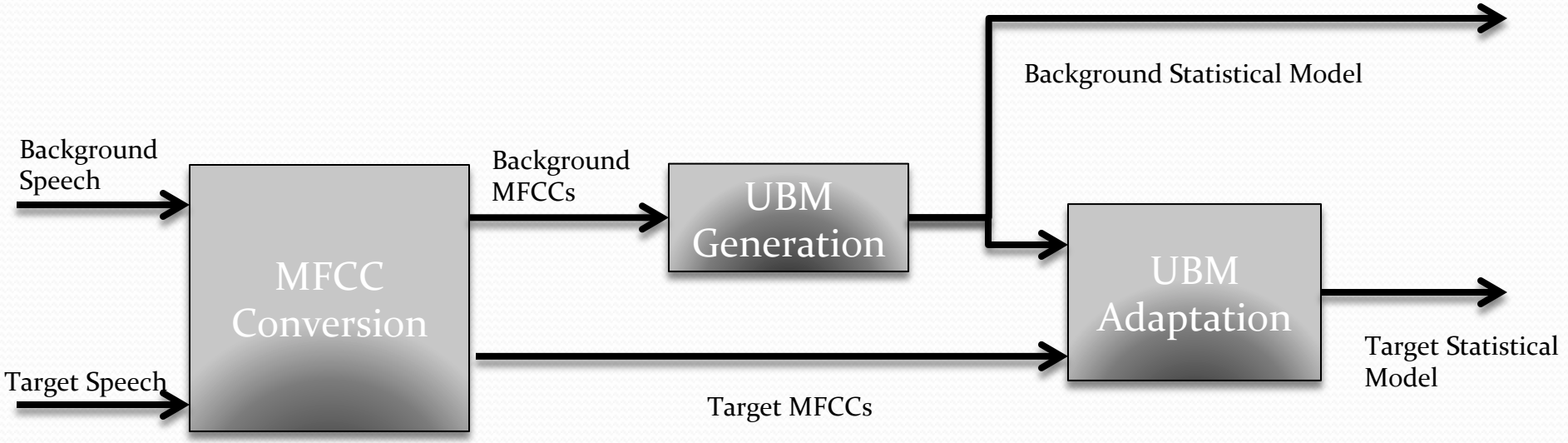
What is Speaker Recognition

- Process of confirming if an unknown speaker is a certain person
- One way to perform this is by using Gaussian Mixture Models (GMM)



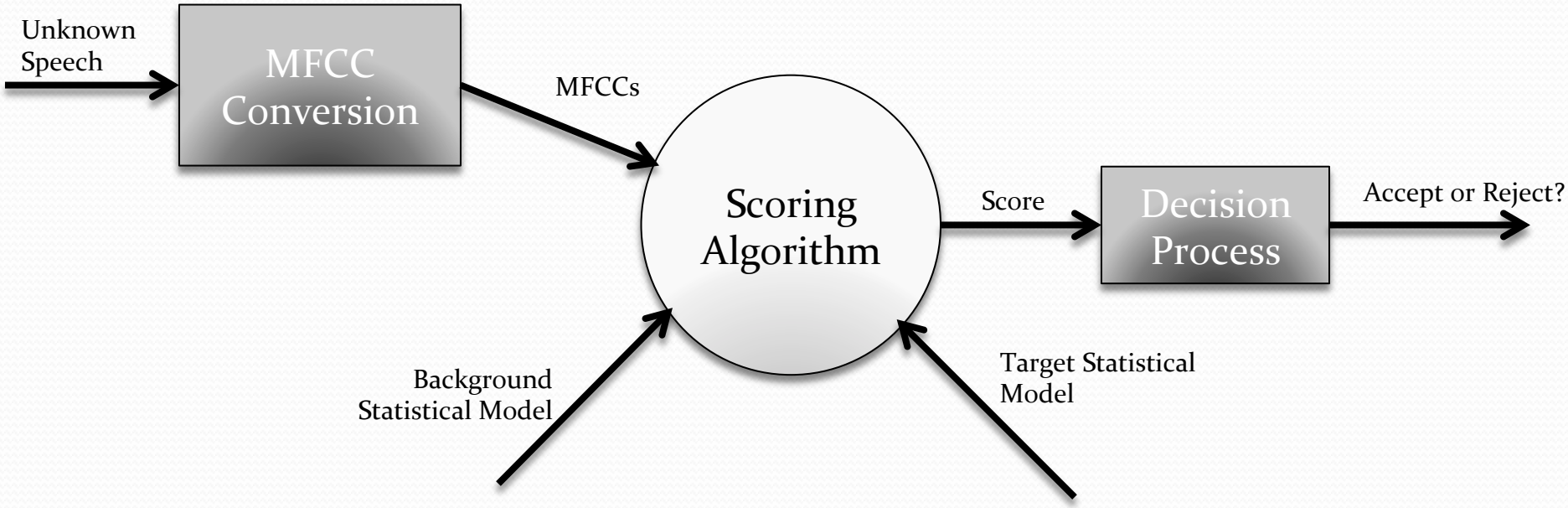
How to Perform Speaker Recognition

- Training Stage

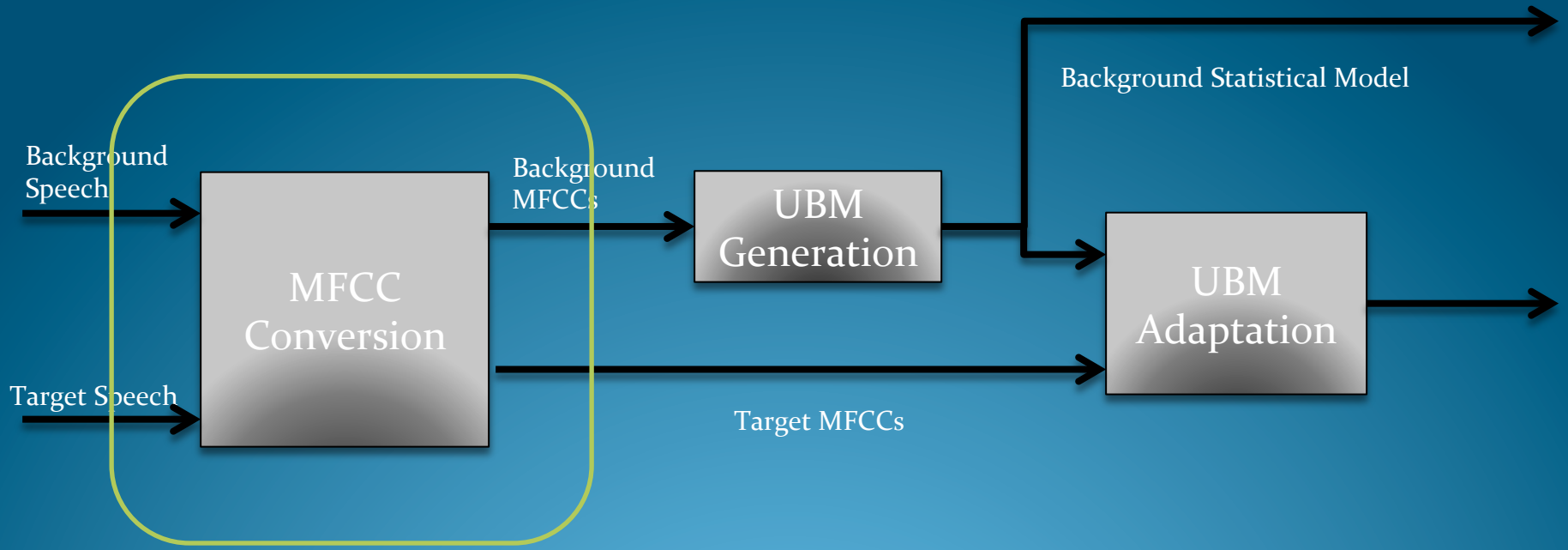


How to Perform Speaker Recognition

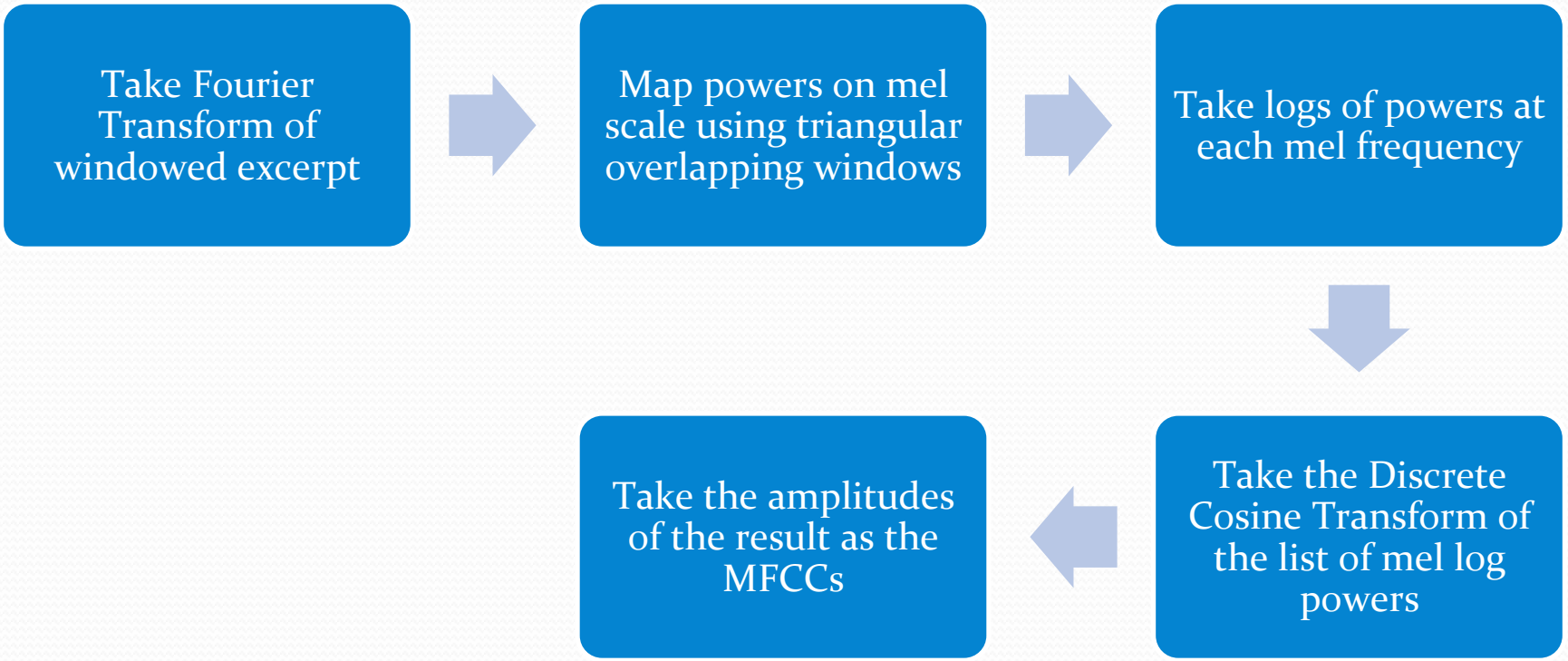
- Testing Stage



1.) Converting to MFCCs

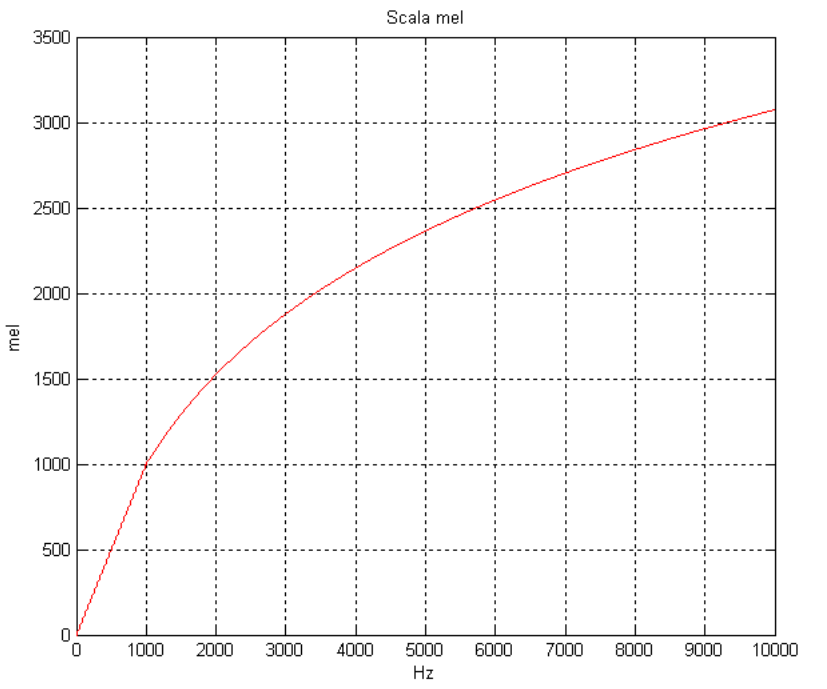


The Process


Take Fourier Transform of windowed excerpt
Map powers on mel scale using triangular overlapping windows
Take logs of powers at each mel frequency
Take the Discrete Cosine Transform of the list of mel log powers
Take the amplitudes of the result as the MFCCs

The Mel Scale

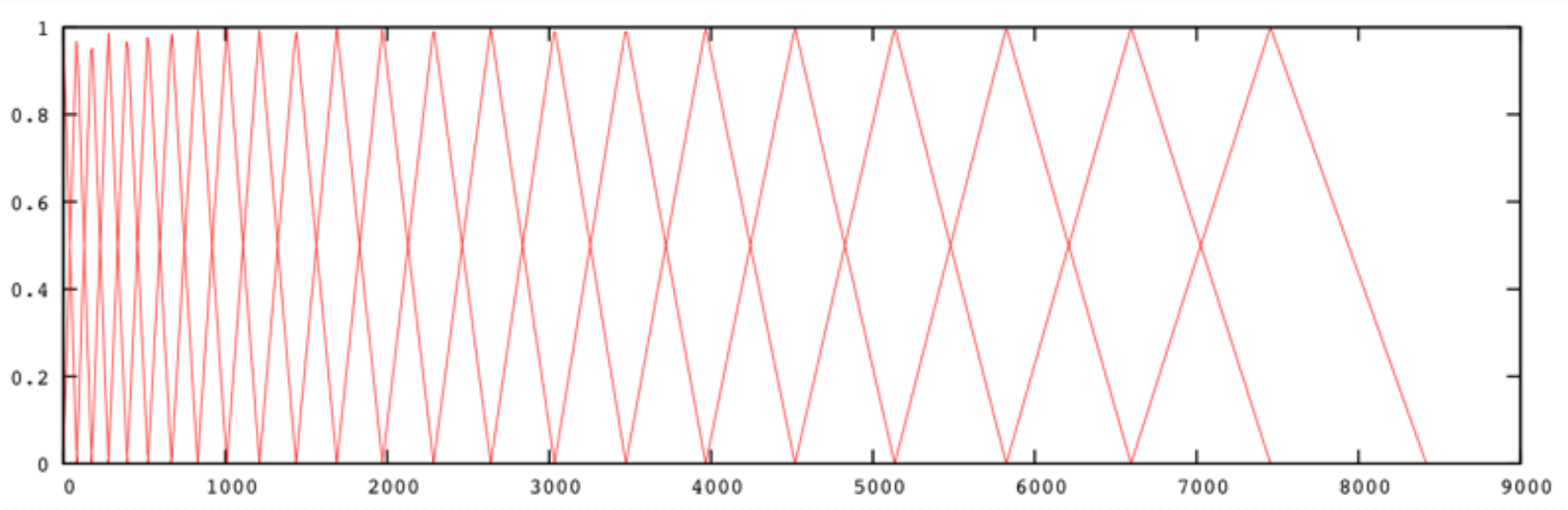
- A nonlinear scale that relates audio frequency to how the human ear hears the frequency.
- Certain frequencies are heard to be about the same pitch by human ears.
- No singular formula because it is so subjective.



$$m = 2595 \log_{10}\left(1 + f/700\right)$$

Triangular Overlapping Windows

- The windows are thought up as filter banks
- The triangles themselves are equally spaced in the mel scale but one applies them in the linear frequency scale
- It can be thought of as a weighted sum for each frequency.



The Discrete Cosine Transform

- Expresses a sequence of data points as a sum of cosine functions oscillating at different frequencies
- Used also for MP3 and JPG compression
- Similar to the Discrete Fourier Transform but using only real numbers

The Discrete Cosine Transform

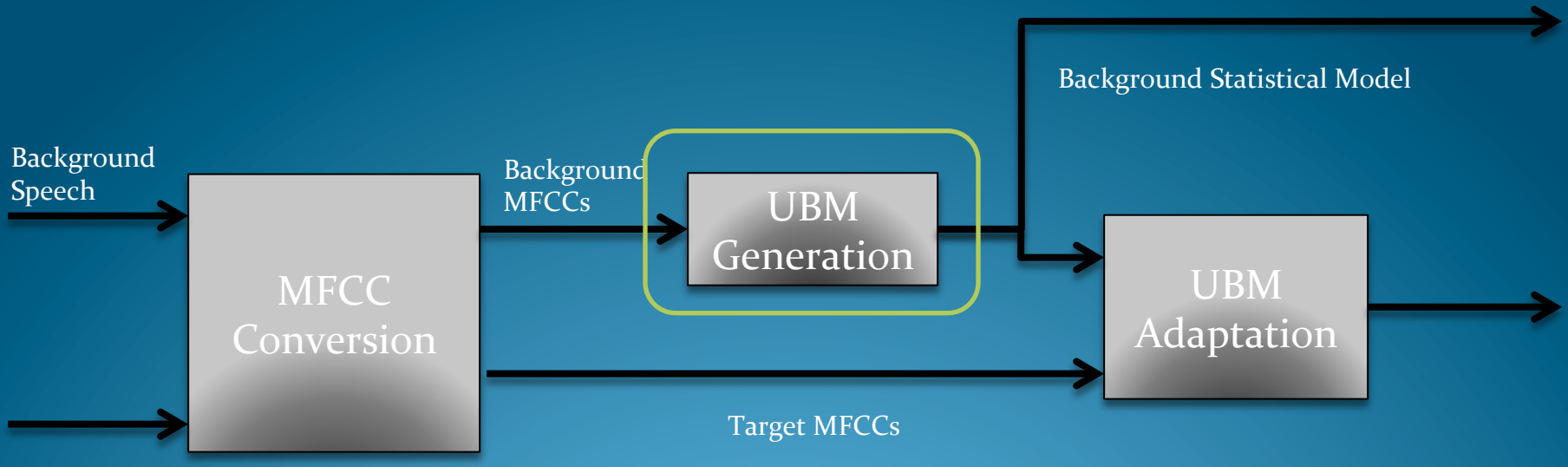
- Multiple forms of the DCT
- The most common one, the DCT-II, is exactly equivalent to a DFT of 4N real inputs of even symmetry where the even-indexed elements are zero.

$$X_k = \sum_{n=0}^{N} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \qquad k=0,\ldots,N-1$$

Deltas and Delta Deltas

- In addition to the raw MFCCs, one also needs to find out the evolution of the tones.
- To find the deltas, one simply finds the difference between the MFCCs dimensions.
- To find the delta deltas, one then finds the differences between the deltas.

2.) Creating the UBM



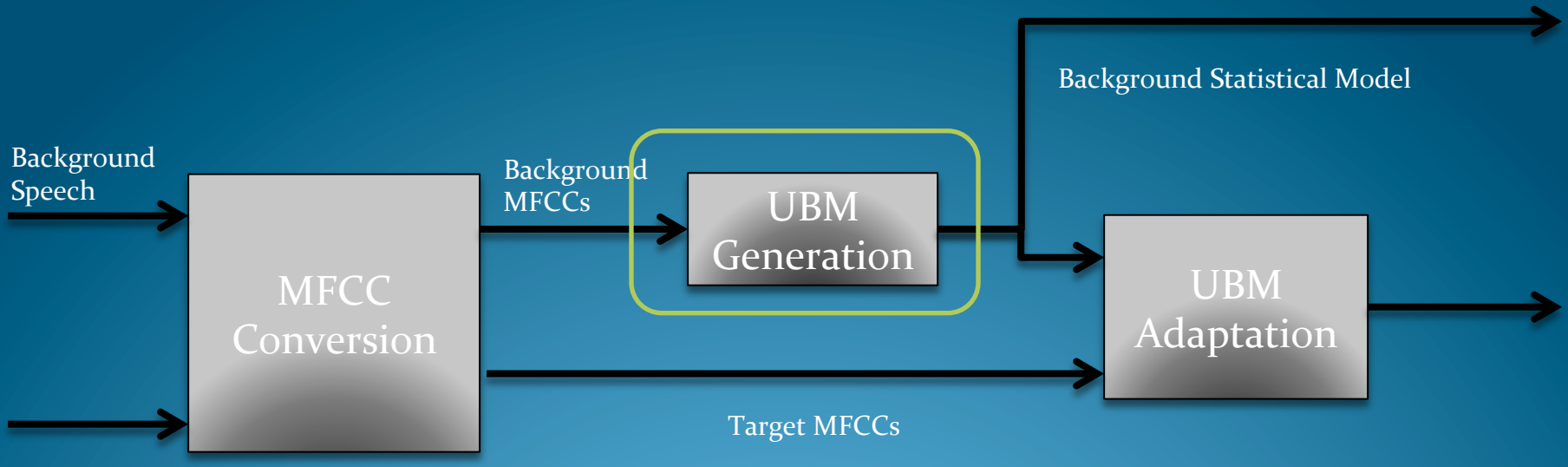
The UBM

- The Universal Background model is the model that corresponds to a generic speaker.
- Create model off of combined speech from many people
- If gender is known, can create a UBM based on the specific gender to get tighter results

The UBM

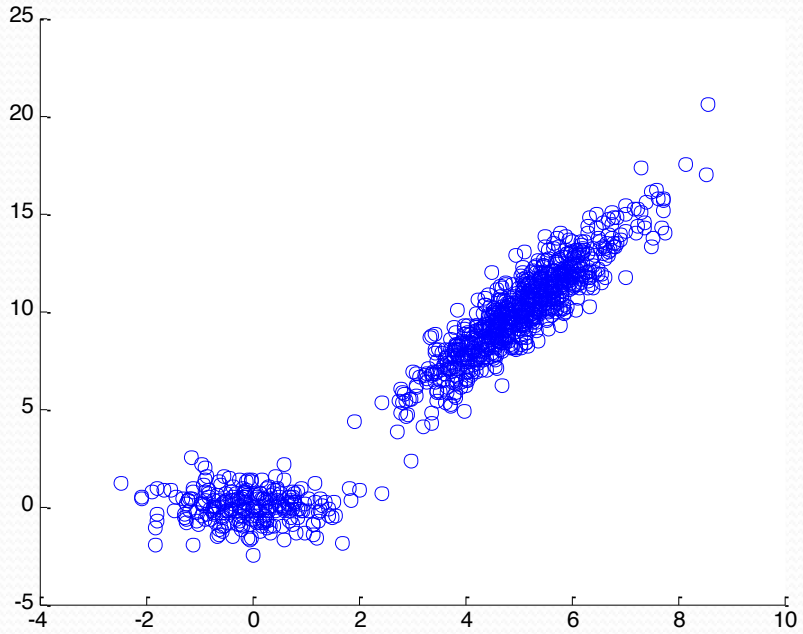
- To form the UBM, one must first generate the MFCCs of many different speakers.
- The simplest method involves outright generating the Gaussian Mixture Model (GMM) parameters based off the MFCCs
- Other models include using K-means clustering first before generating the GMM parameters

Gaussian Mixture Model



How can we classify data

- Supposed we have data shown to the right
- How can we assign a probability distribution to this data to show how it was created?



The Gaussian Mixture Model

- Weighted sum of M component Gaussian densities given by the equation

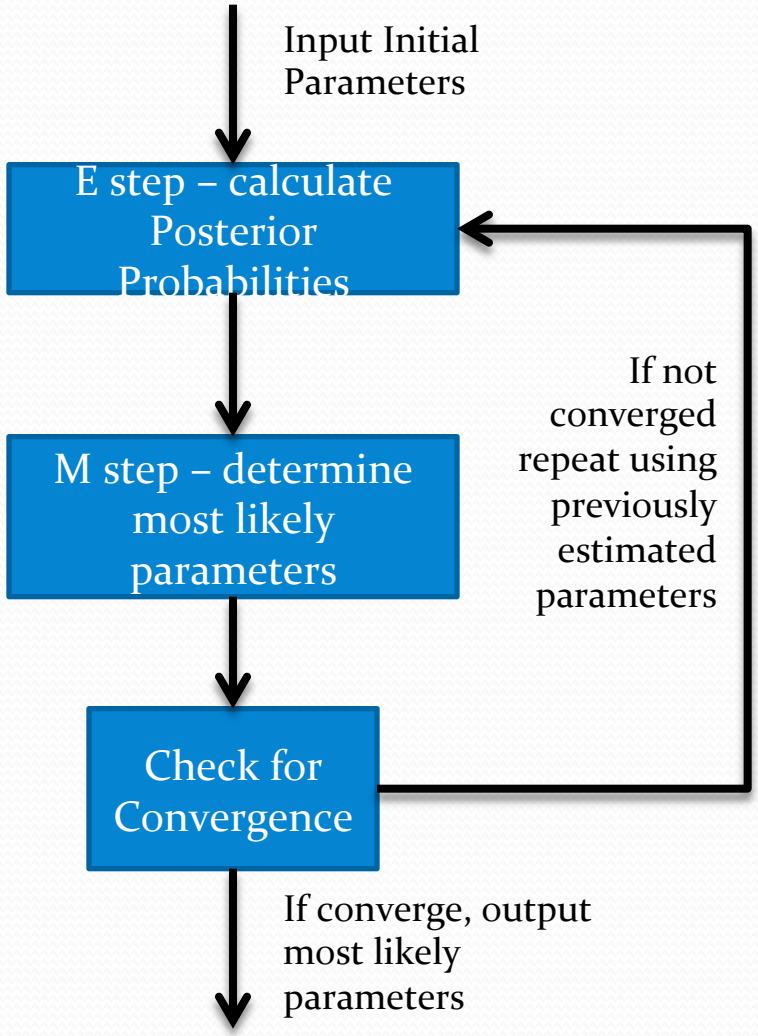
$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \, g(\mathbf{x}|\boldsymbol{\mu}_i, \Sigma_i)$$

- Where

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)}$$

- Generate using Expectation Maximization (EM)

Expectation Maximization



The Algorithm

- Calculate the posterior probabilities of all the data points for each class

$$h_i^{(j)}(t) = w_i^{(j)} g(\mathbf{x}_t \mid \boldsymbol{\mu}_i^{(j)}, \Sigma_i^{(j)}) / \sum_{s=1}^{n} w_s^{(j)} g(\mathbf{x}_t \mid \boldsymbol{\mu}_s^{(j)}, \Sigma_s^{(j)})$$

$$s_i^{(j)} = \sum_{t=1}^{N} h_i^{(j)}(t)$$

The Algorithm

- Calculate the parameters for the next iteration

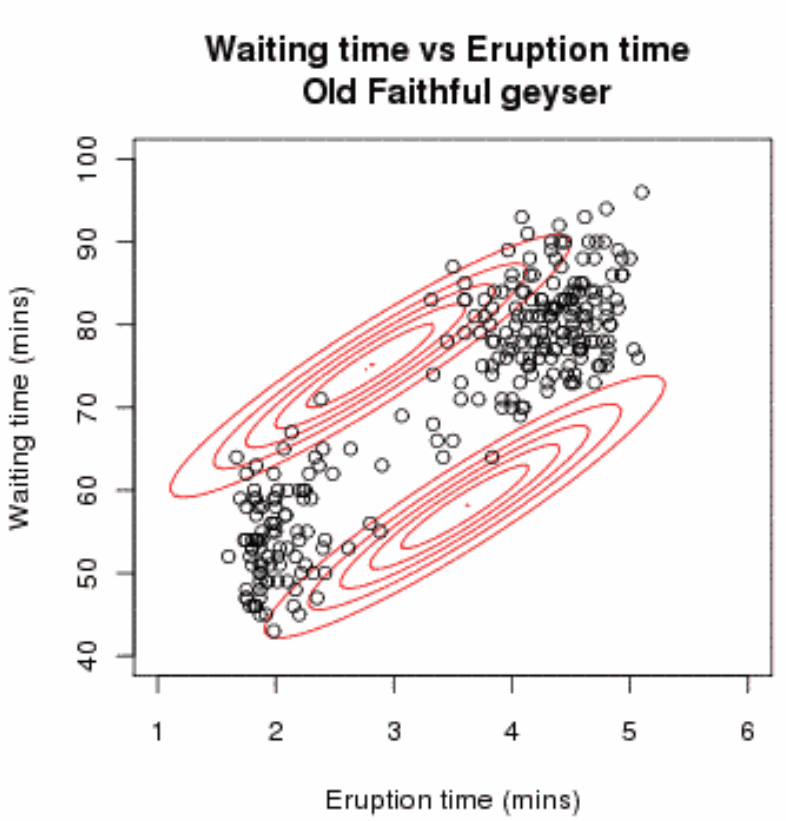
$$w_i^{(j+1)} = 1/N \, s_i^{(j)}$$

$$\boldsymbol{\mu}_i^{(j+1)} = \sum_{t=1}^{N} h_i^{(j)}(t) \, \mathbf{x}_t / s_i^{(j)}$$

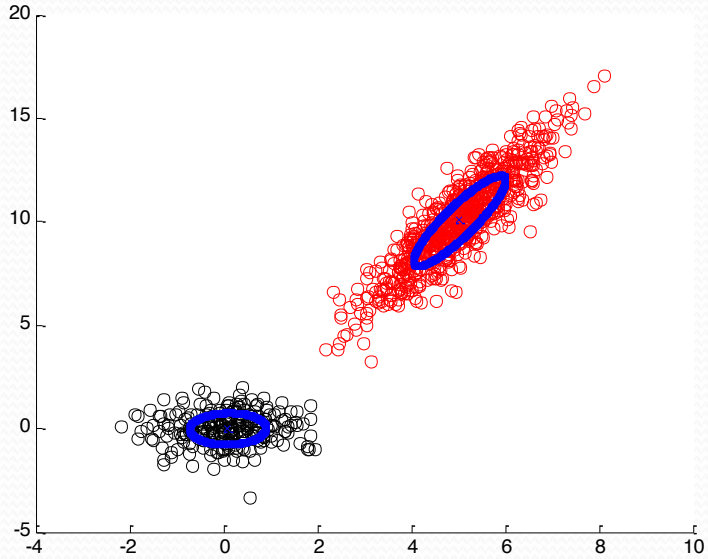
$$\Sigma_i^{(j+1)} = \sum_{t=1}^{N} h_i^{(j)}(t) [\mathbf{x}_t - \boldsymbol{\mu}_i^{(j)}][\mathbf{x}_t - \boldsymbol{\mu}_i^{(j)}]^{\mathrm{T}} / s_i^{(j)}$$

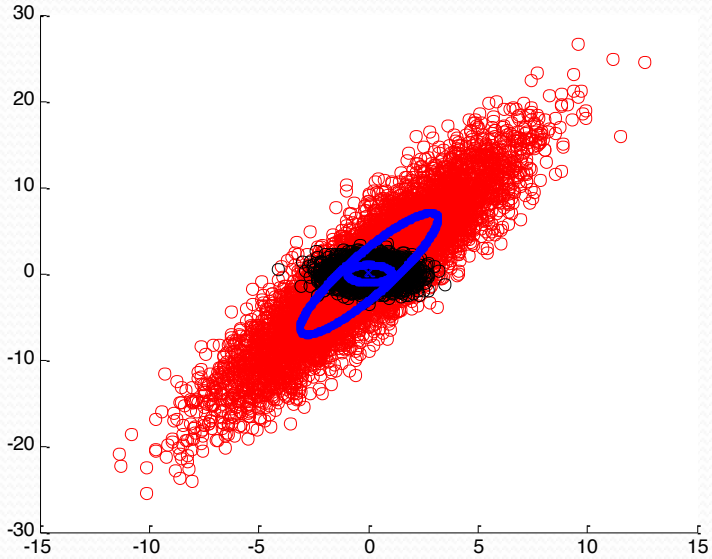
Examples of Generic GMM adaptation

- To the right is an example of the GMM algorithm working on the Old Faithful dataset



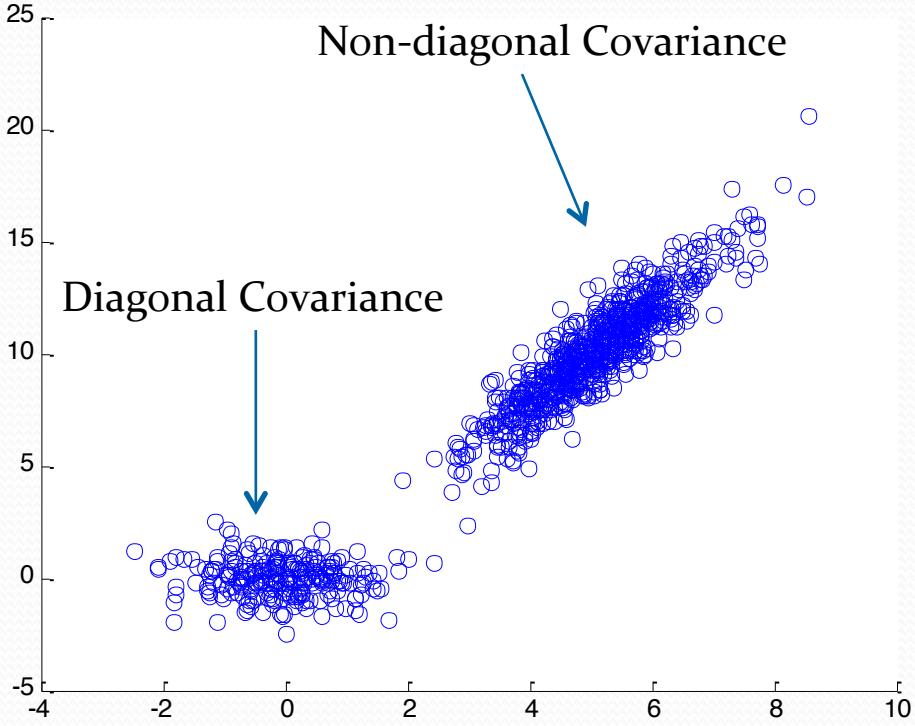
Some Created Data





The Covariance Matrix

- One does not typically care for the off diagonal terms of the covariance matricies.
- Calculations become intensive if a full covariance matrix is used. In some cases, they hurt the error rate



Alternate Representation of the Code

- Problem is that it requires a lot of memory and processing power to compute and store the pdfs of every point and every mixture.
- Large volumes of data, high dimensionality, and many mixture coefficients make running the process hard using the standard form

Alternate Representation of the Code

- One can first take the log of the pdf to form the following:

$$L_{p,j} = \log w_j - 1/2 \log |2\pi\Sigma_j| - 1/2 \, (\mathbf{x}_t - \boldsymbol{\mu}_j)^{\mathrm{T}} \Sigma_j^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_j) = \log w_j - 1/2 \log |2\pi\Sigma_j| - 1/2 \sum_{i=1}^{D} \sigma_{i,j}^{-1} \mu_{i,j}^2 + \sum_{i=1}^{D} \mu_{i,j} \sigma_{i,j}^{-1} x_{i,t} - 1/2 \sum_{i=1}^{D} \sigma_{i,j}^{-1} x_{i,t}^2$$

Alternate Representation of the Code

- At the start of each iteration, one can save time by precomputing the following:

$$A = \log w_j - 1/2 \log |2\pi\Sigma_j| - 1/2 \sum_{i=1}^{D} \sigma_{i,j}^{-1} \mu_{i,j}^2$$

- In addition, one can queue up the coefficients as the following

$$B_i = \mu_{i,j} \sigma_{i,j}^{-1}$$

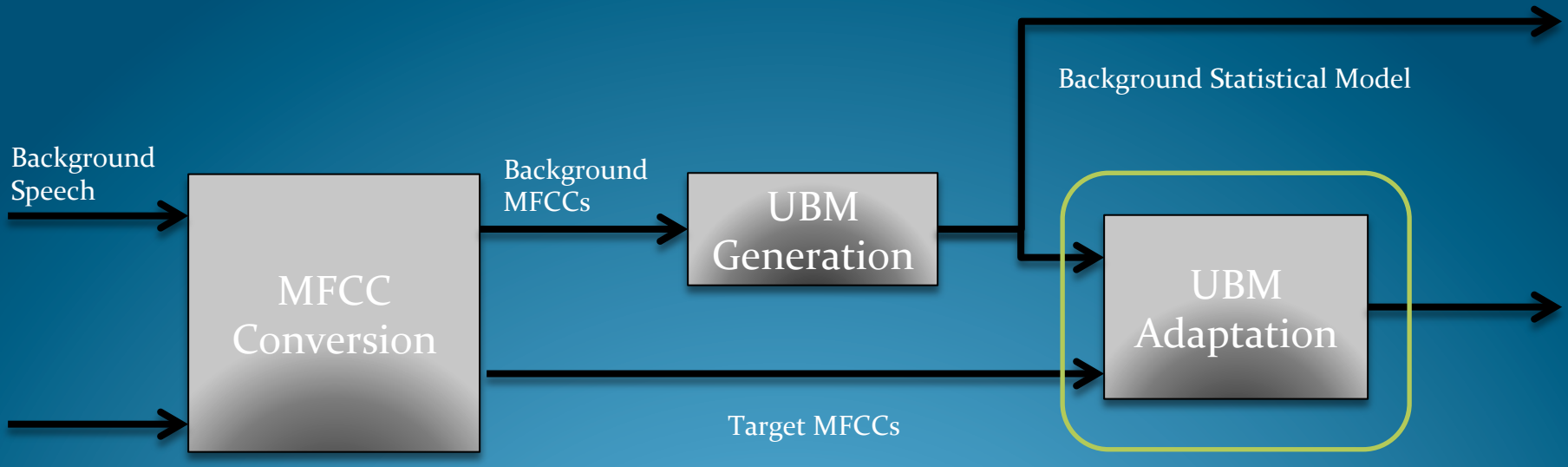
$$C_i = -1/2 \sigma_{i,j}^{-1}$$

Alternate Representation of the Code

- Compute x_i^2 at the beginning of the code.
  - Beneficial to actually accept the x_i^2 as an input parameter.

Cycle through Points	Cycle through Mixtures	Calculate Probability of Existence in Each Mixture
	Cycle through Mixtures	Calculate Posterior Probabilities and Rolling Means, Covariances, and Weights
Cycle through Mixtures		Finalize Means, Covariances, and Weights

3.) UBM Adaptation



UBM Adaptation

- Typically our target speaker does not provide us with nearly as much data as what we want.
- Because of that, creating a GMM from scratch will not produce a very accurate model
- The UBM parameters can be adjusted to the parameters of the target speaker

The Algorithm

- Like with the GMM training, the first step is to compute the posterior probabilities

$$h_i^{(t)} = \frac{w_i \, g(\mathbf{x}^t \mid \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{s=1}^{n} w_s \, g(\mathbf{x}^t \mid \boldsymbol{\mu}_s, \Sigma_s)}$$

The Algorithm

- From there, one calculates the sufficient statistics for the means, covariances, and weights

$$n_i = \sum_{t=1}^{T} h_i(t)$$

$$E_i(\mathbf{x}_t) = 1/n_i \sum_{t=1}^{T} h_i(t)\, \mathbf{x}_t$$

$$E_i(\mathbf{x}_t^2) = 1/n_i \sum_{t=1}^{T} h_i(t)\, \mathbf{x}_t^2$$

The Algorithm

- Using these sufficient statistics and the old background model sufficient statistics, the new estimate of the means, covariances, and weights can be produced

$$w_i^{(j+1)} = [\alpha_i n_i / T + (1-\alpha_i) w_i^{(j)}]\gamma$$

$$\mu_i^{(j+1)} = \alpha_i E_i(\mathbf{x}) + (1-\alpha_i)\mu_i^{(j)}$$

$$\Sigma_i^{(j+1)} = \alpha_i E_i(\mathbf{x}^2) + (1-\alpha_i)(\Sigma_i^{(j)} + \mu_i^{(j)2}) - \mu_i^{(j+1)2}$$

- Where γ is a scale factor so the weights sum to 1 and α_i is defined as the following

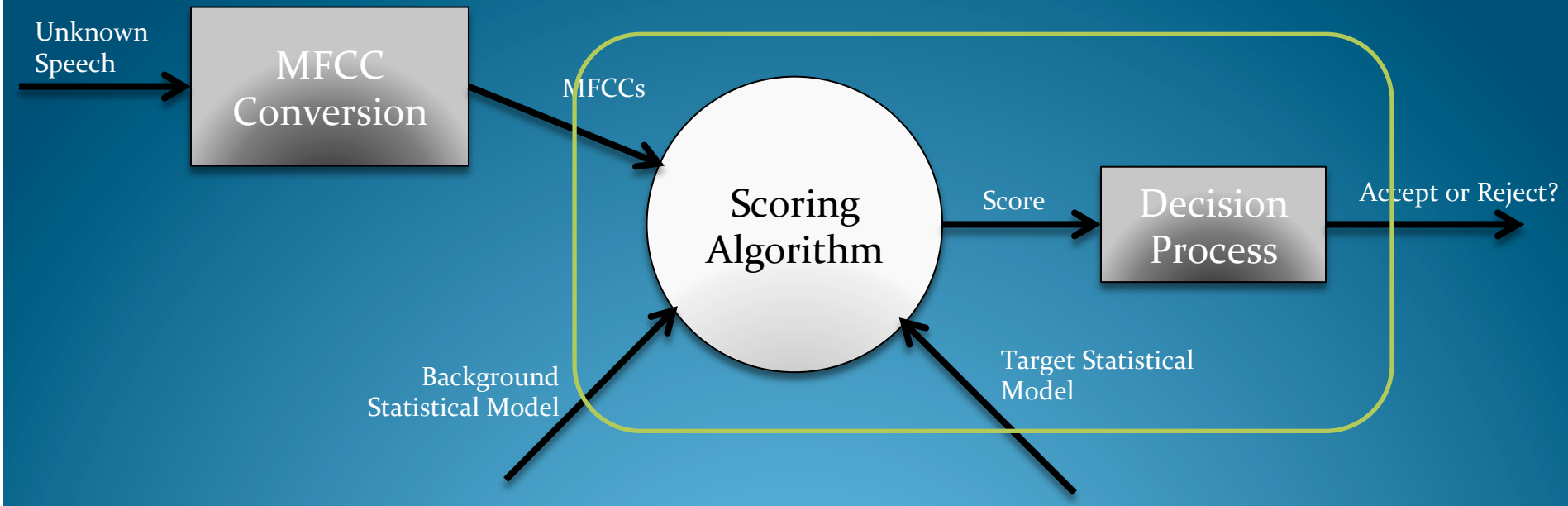
$$\alpha_i = n_i / n_i + r$$

- Where r is a fixed relevance parameter

UBM Adaptation

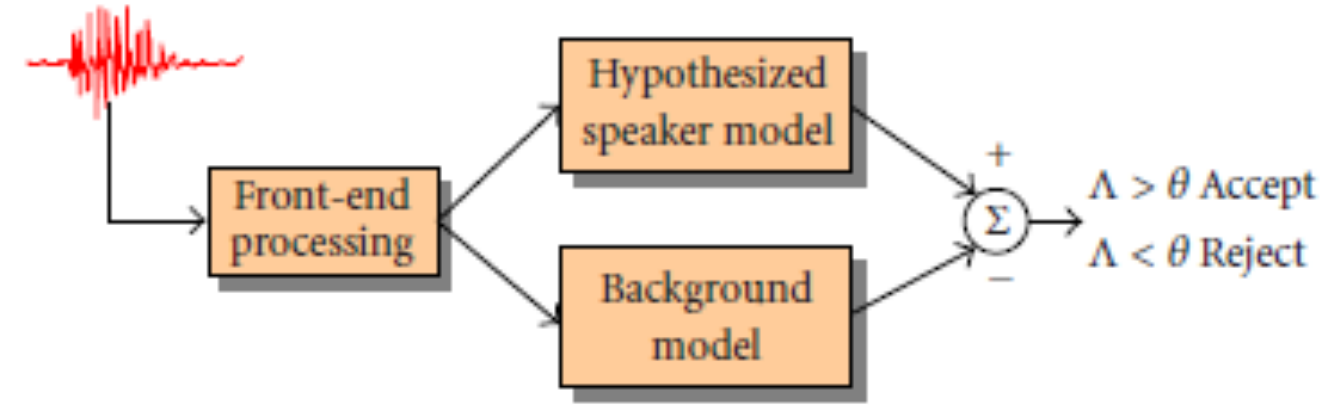
- With data that has low probabilistic count of new data, the new data is deemphasized and the old data is particularly emphasized.
  - The reverse is true when the data has high probabilistic count of new data
- Because covariances and weights aren't primary parameters, adapting them with such a small dataset isn't a very good idea.
  - Most of the times, only the means are adapted

4.) Scoring


Unknown Speech
MFCC Conversion
MFCCs
Scoring Algorithm
Score
Decision Process
Accept or Reject?
Background Statistical Model
Target Statistical Model

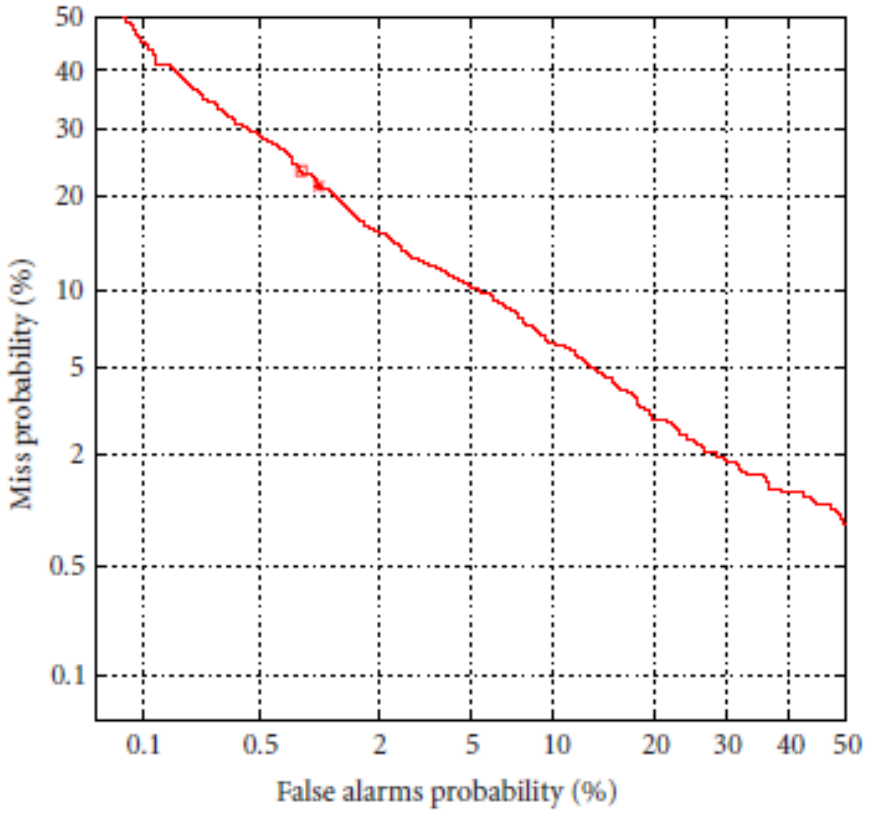
Scoring

- A simple log likelihood test is performed using the sufficient statistics of the UBM and the adapted UBM when one gets new data



DET Curves

- One cannot just choose a specific acceptance value.
- For each threshold for acceptance, a certain false acceptance and false rejection rate is generated



DET Curves

- One way to accept, is to find the values that give an EER (Equal Error Rate)
- If one wants to accept more or reject more, one can solve an optimization problem
- Cost is defined as

$$C = C_{fa} P_{fa} + C_{fr} P_{fr}$$

Conclusion

Conclusion

- To begin a speaker recognition system, one must first have the computer train up a background model and adapt the background model to every speaker.
  - This model can be saved in memory and does not have to be recomputed every time
- Once a person with an unknown identity is speaking, the computer can score it against a background model to see if it's the desired target speaker or not