Speaker Verification

- likelihood ratio detector

\[
\frac{p(\text{speech}|\text{speaker})}{p(\text{speech}|\text{speaker})} = \frac{p(X|\lambda_{\text{speaker}})}{p(X|\lambda_{\text{speaker}})} \begin{cases} 
\geq \theta & \text{accept speaker} \\
< \theta & \text{reject speaker}
\end{cases}
\]

- $\lambda_{\text{speaker}}$ is well defined, as it is estimated using training speech from the speaker

- $\lambda_{\text{speaker}}$ is not so well defined, as it represents the entire space of alternatives to the speaker

  * two possible solutions: create a set of models from other speakers OR pool speech from several speakers and train a single model

  * the second solution is better, as it speaker independent

- gaussian mixture models (GMMs)

\[
p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x)
\]

\[
p_i(x) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\{-1/2(x - \mu_i)'(\Sigma_i)^{-1}(x - \mu_i)\}
\]

\[
\sum_{i=1}^{M} w_i = 1
\]

- $\lambda = \{w_i, \mu_i, \Sigma_i\}$

  GMM parameters refined using the EM algorithm

- universal background model (UBM)

  - two solutions for estimation of the final UBM: pool all the data to train the UBM using the EM algorithm OR train individual UBMs over sub-populations and then pool the sub-populations models

  - the second solution is better, as it allows the use of unbalanced data for training
adaptation of the speaker model

- derive speaker model by adapting the parameters of the UBM using the speaker’s training speech and by Bayesian adaptation (maximum a posteriori estimation)

- steps for speaker model adaptation:
  * compute estimates of sufficient statistics of the speaker’s training data for each mixture in the UBM (identical to expectation step of the EM algorithm)
  * adapt UBM to the speaker: mixtures with high data counts use new statistics (speaker), mixtures with low data counts use old statistics (UBM)

- equations for speaker model adaptation:
  * probabilistic alignment of training vectors into UBM:
    \[
    Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}
    \]
  * new sufficient statistics (step 1):
    \[
    n_i = \sum_{t=1}^{T} Pr(i|x_t)
    \]
    \[
    E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t)x_t
    \]
    \[
    E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t)x_t^2
    \]
  * update old sufficient statistics (step 2):
    \[
    \hat{w}_i = [\alpha^w_t n_t \gamma + (1 - \alpha^w_t)w_i] \gamma
    \]
    \[
    \hat{\mu}_i = \alpha^m_t E_i(x) + (1 - \alpha^m_t)\mu_i
    \]
    \[
    \hat{\sigma}_i^2 = \alpha^v_t E_i(x^2) + (1 - \alpha^v_t)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2
    \]

  \[
  \alpha^w_t, \alpha^m_t, \alpha^v_t: \text{balance old and new statistics}
  \]
  \[
  \gamma : \sum \hat{w}_i = 1
  \]