- Maximum Likelihood Estimation (MLE)

⇒ choose model parameter λ to make likelihood $P_\lambda(Y|W)$ as large as possible

($Y \rightarrow$ acoustic signal, $W \rightarrow$ word sequence)

$$F_{ML}(\lambda) = \sum_u P_\lambda(Y_u|W_u)$$

- Maximum Mutual Information Estimation (MMIE)

⇒ choose model parameter λ to make mutual information $I_\lambda(W,Y)$ as large as possible.

$$I_\lambda(W,Y) = \sum_{W,Y} P(W,Y) \log\left(\frac{P_\lambda(W,Y)}{P_\lambda(W)P_\lambda(Y)}\right)$$

$$\Rightarrow F_{MMIE}(\lambda) = \sum_u \log\left(\frac{P_\lambda(W,Y)}{P_\lambda(W)P_\lambda(Y)}\right)$$

$$= \sum_u \log\left(\left(\frac{P_\lambda(W,Y)}{P_\lambda(W)}\right)\frac{1}{P_\lambda(Y)}\right) = \sum_u \left(\log P_\lambda(Y|W) - \log P_\lambda(Y)\right)$$

($\log P_\lambda(Y|W)$ → MLE)

⇒ The difference btw MLE, MMIE = $-\log P_\lambda(Y)$

(ignore outermost Σ)

$$\frac{\partial F_{MMIE}(\lambda)}{\partial\lambda} \doteq \frac{\partial\{\log P_\lambda(Y|W) - \log P_\lambda(Y)\}}{\partial\lambda} = \frac{\partial \log P_\lambda(Y|W)}{\partial\lambda} - \frac{\partial \log P_\lambda(Y)}{\partial\lambda}$$

$$= \frac{\left\{\frac{\partial P_\lambda(Y|W)}{\partial\lambda}\right\}}{P_\lambda(Y|W)} - \frac{\left\{\frac{\partial P_\lambda(Y)}{\partial\lambda}\right\}}{P_\lambda(Y)} = \frac{\left\{\frac{\partial P_\lambda(Y|W)}{\partial\lambda}\right\}}{P_\lambda(Y|W)} - \frac{\left\{\frac{\partial\left(\sum_{\tilde{W}} P(Y|\tilde{W})P(\tilde{W})\right)}{\partial\lambda}\right\}}{P_\lambda(Y)}$$

$$= \frac{\left\{\frac{\partial P_\lambda(Y|W)}{\partial\lambda}\right\}}{P_\lambda(Y|W)} - \frac{\left\{\frac{\partial P_\lambda(Y|W)}{\partial\lambda}\right\}P(W) + \sum_{\tilde{W}\neq W}\left\{\frac{\partial P_\lambda(Y|\tilde{W})}{\partial\lambda}\right\}P(\tilde{W})}{P_\lambda(Y)}$$

$$= \left\{\frac{\partial P_\lambda(Y|W)}{\partial\lambda}\right\}\left\{\frac{1}{P_\lambda(Y|W)} - \frac{P(W)}{P_\lambda(Y)}\right\} - \sum_{\tilde{W}\neq W}\frac{\left\{\frac{\partial P_\lambda(Y|\tilde{W})}{\partial\lambda}\right\}P(\tilde{W})}{P_\lambda(Y)}$$

↓ derivative of MLE

↓ Subtract direction $\frac{\partial P_\lambda(Y|\tilde{W})}{\partial\lambda}$ for incorrect word sequence $\tilde{W} \neq W$

→ $F_{mmie}(\lambda) = \sum_u \log\left(\frac{P_\lambda(W,Y)}{P_\lambda(W)P_\lambda(Y)}\right) = \sum_u \left\{\log\frac{P_\lambda(W,Y)}{P_\lambda(Y)} - \log P_\lambda(W)\right\}$

$= \sum_u \left\{\log\frac{P_\lambda(Y|W)^k P(W)}{\sum_w P_\lambda(Y|W)^k P(W)} - \log P_\lambda(W)\right\}$

(k = scaling fudge factor)

- Boosted MMI

⇒ boost the likelihood of the sentences have ↑ errors, generating ↑ confusable data.

→ $F_{mmie}(\lambda) = \sum_u \left\{\log\frac{P_\lambda(Y|W)^k P(W)}{\sum_w P_\lambda(Y|W)^k P(W)\, e^{-bA(W,W_u)}}\right\}$

(b = boosting factor, $A(W,W_u)$ = accuracy of sentence W given reference W_u)

- Minimum Phone Error (MPE), minimum Bayes Risk (s-MBR) (state-level)

⇒ minimize the error corresponding to different granularity of labels (phone / state)

→ $F_{mpe}(\lambda) = \sum_u \left\{\log\frac{\sum_w P_\lambda(Y|W)^k P(W) A(W,W_u)}{\sum_w P_\lambda(Y|W)^k P(W)}\right\}$

($A(W,W_u)$ = phone accuracy of W given reference W_u)

→ $F_{smbr}(\lambda)$ ⇒ $A(W,W_u)$ = state-level accuracy