Discriminative Training Approaches for Continuous Speech Recognition

Berlin Chen, Jen-Wei Kuo, Shih-Hung Liu

Speech Lab

Graduate Institute of Computer Science & Information Engineering

National Taiwan Normal University

Outline

- Statistical Speech Recognition
- Overall Expected Risk Decision/Estimation
- Maximum Likelihood (ML) Estimation
- Maximum Mutual Information (MMI) Estimation
- Minimum Phone Error (MPE) Estimation
- Preliminary Experimental Results
- Conclusions

Statistical Speech Recognition (1/2)

- The task of the speech recognition is to determine the identity of an given observation sequence O by assigning the recognized word sequence W to it

- The decision is to find the identity with maximum posterior (MAP) probability $p(W \mid O)$
  - The so-called Bayes decision (or minimum-error-rate) rule

$$W^* = \arg\max_W p(W \mid O) = \arg\max_W \frac{p(O \mid W)p(W)}{p(O)} \approx \arg\max_W p(O \mid W)p(W)$$

Statistical Speech Recognition (2/2)

- The probabilities needed for the Bayes decision rule are not given in practice and have to be estimated from training dataset
- Moreover, as the true families of distributions to which $p(W)$ (typically assumed Multinomial) and $p(O|W)$ (typically assumed Gaussian) belong are not known
  - A certain parametric representation of these distributions is needed
  - In this presentation, language model $p(W)$ is assumed to be given in advance while acoustic model $p(O|W)$ is needed to be estimated
    - HMMs (hidden Markov models) are widely adopted for acoustic modeling

Expected Risk

- Let $\mathbf{W}^r = \{W_1, W_2, W_3, \cdots\}$ be a finite set of various possible word sequences for a given observation utterance O_r
  - Assume that the true word sequence W_r is also in $\mathbf{W}^r$

- Let $\alpha(O_r) = W_j$ be the action of classifying a given observation sequence O_r to a word sequence $W_j \in \mathbf{W}^r$
  - Let $l(W_n, W_j)$ be the loss incurred when we take such an action (and the true word sequence W_r is just W_n $(W_n = W_r)$)

- Therefore, the (expected) risk for a specific action $\alpha(O_r) = W_j$

$$R(\alpha(O_r) == W_j \mid O_r) = \sum_{W_n \in \mathbf{W}^r} l(W_n, W_j) P(W_n \mid O_r)$$

Decoding: Minimum Expected Risk (1/2)

- In speech recognition, we can take the action with the minimum (expected) risk

$$\arg\min_{W_j} R(\alpha(O_r) == W_j \mid O_r) = \arg\min_{W_j} \sum_{W_n \in \mathbf{W}^r} l(W_n, W_j) P(W_n \mid O_r)$$

- If zero-one loss function is adopted (string-level error)

$$l(W_n, W_j) = \begin{cases} 0 & \text{if} \quad W_n = W_j \\ 1 & \text{if} \quad W_n \neq W_j \end{cases}$$

  - Then

$$\arg\min_{W_j} R(\alpha(O_r) == W_j \mid O_r) = \arg\min_{W_j} \sum_{W_n \in \mathbf{W}^r, W_n \neq W_j} P(W_n \mid O_r)$$

$$= \arg\min_{W_j} \left[1 - P(W_j \mid O_r)\right]$$

Decoding: Minimum Expected Risk (2/2)

– Thus,

$$\arg\min_{W_j} R(\alpha(O_r) == W_j \mid O_r) = \arg\max_{W_j} P(W_j \mid O_r)$$

• Select the word sequence with maximum posterior probability (MAP decoding) (cf. p.3)

- The string editing or Levenshtein distance also can be accounted for the loss function
  – Take individual word errors into consideration
  – E.g., Minimum Bayes Risk (MBR) search/decoding [V. Goel et al. 2004]

Training: Minimum Overall Expected Risk (1/2)

- In training, we should minimize the overall (expected) loss of the actions $\alpha(O_r) = W_r$ of the training utterances O_r
  - W_r is the true word sequence of O_r

$$R_{overall} = \int R(\alpha(O_r) == W_r \mid O_r) p(O_r) dO_r$$

  - The integral extends over the whole observation sequence space
- However, when a limited number R of training observation sequences are available, the overall risk can be approximated by

$$R_{overall} \approx \sum_{r=1}^{R} R(\alpha(O_r) == W_r \mid O_r) p(O_r)$$

$$= \sum_{r=1}^{R} \sum_{W_n \in \mathbf{W}^r} l(W_n, W_r) P(W_n \mid O_r) p(O_r)$$

Training: Minimum Overall Expected Risk (2/2)

- Assume $p(O_r)$ to be uniform
  - The overall risk can be further expressed as

$$R_{overall} = \sum_{r=1}^{R} \sum_{W_n \in \mathbf{W}^r} l(W_n, W_r) P(W_n \mid O_r)$$

- If zero-one loss function is adopted

$$l(W_n, W_r) = \begin{cases} 0 & \text{if} \quad W_n = W_r \\ 1 & \text{if} \quad W_n \neq W_r \end{cases}$$

  - Then

$$R_{Overall} = \Sigma_r \sum_{W_n \in \mathbf{W}^r, W_n \neq W_r} P(W_n | O_r) = \Sigma_r \left[1 - P(W_r | O_r)\right]$$

Training: Maximum Likelihood (1/3)

- The objective function of Maximum Likelihood (ML) estimation can be obtained if Jensen Inequality is further applied

$$R_{Overall} = \sum_r \left[1 - P\left(W_r \middle| O_r\right)\right] \leq \sum_r -\log P\left(W_r \middle| O_r\right)$$

minimize the upper bound

$$x - 1 \geq \log x \Rightarrow 1 - x \leq -\log x$$

  - Find a new parameter set that minimizes the overall expected risk is equivalent to those that maximizes the overall log likelihood of all training utterances

maximize the low bound

$$\lambda_{ML} = \arg\max_{\lambda} \sum_r \log P\left(W_r \middle| O_r\right) = \arg\max_{\lambda} \sum_r \log \frac{P_\lambda\left(O_r \middle| W_r\right) P\left(W_r\right)}{P\left(O_r\right)}$$

$$\approx \arg\max_{\lambda} \sum_r \log P_\lambda\left(O_r \middle| W_r\right)$$

$$\Rightarrow F_{ML}(\lambda) = \sum_r \log P_\lambda\left(O_r \middle| W_r\right)$$

Training: Maximum Likelihood (2/3)

- The objective function can be maximized by adjusting the parameter set, with the EM algorithm and a specific auxiliary function (or the Baum-Welch algorithm)

$$F_{ML}(\lambda) = \sum_r \log P_\lambda(O_r | W_r)$$

$$F_{ML}(\lambda) = Q(\lambda, \bar{\lambda}) - H(\lambda, \bar{\lambda})$$
$$F_{ML}(\lambda) - F_{ML}(\bar{\lambda}) \geq Q(\lambda, \bar{\lambda}) - Q(\bar{\lambda}, \bar{\lambda})$$

  - E.g., update formulas for Gaussians

$$\mu_{qm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{qm}^{r}(t)O_r(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{qm}^{r}(t)}$$

$$\sigma_{qm}^2 = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{qm}^{r}(t)O_r(t)^2}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{qm}^{r}(t)} - \mu_{qm}^2$$

Training: Maximum Likelihood (3/3)

- On the other hand, the discriminative training approaches attempt to optimize the correctness of the model set by formulating an objective function that in some way penalizes the model parameters that are liable to confuse correct and incorrect answers

Training: Maximum Mutual Information (1/4)

- The objective function can be defined as the sum of the pointwise mutual information of all training utterances and their associated true word sequences

$$F_{MMI}(\lambda) = \sum_r \mathrm{MI}(O_r, W_r)$$

$$= \sum_r \log \frac{P(O_r, W_r)}{P(O_r)P(W_r)} = \sum_r \log \frac{P_\lambda(O_r|W_r)}{P(O_r)}$$

$$\approx \sum_r \log \frac{P_\lambda(O_r|W_r)}{\sum_{W_k \in \mathbf{W}^r} P_\lambda(O_r|W_k)P(W_k)}$$

  - A kind of rational functions
- The maximum mutual information (MMI) estimation tries to find a new parameter set (λ) that maximizes the above objective function

Training: Maximum Mutual Information (2/4)

- An alternative derivation based on the overall expected risk criterion (cf. p.10)
  - zero-one loss function

$$R_{Overall} = \sum_r \left[1 - P\left(W_r \middle| O_r\right)\right]$$
$$\leq \sum_r -\log P\left(W_r \middle| O_r\right)$$

  - Which is equivalent to the maximization of the overall log likelihood of training utterances

$$\lambda_{MMI} = \arg\max_{\lambda} \sum_r \log P\left(W_r \middle| O_r\right) = \arg\max_{\lambda} \sum_r \log \frac{P_\lambda\left(O_r \middle| W_r\right) P\left(W_r\right)}{P\left(O_r\right)}$$

$$\approx \arg\max_{\lambda} \sum_r \log \frac{P_\lambda\left(O_r \middle| W_r\right)}{\sum_{W_k \in \mathbf{W}^r} P_\lambda\left(O_r \middle| W_k\right) P\left(W_k\right)}$$

$$\Rightarrow F_{MMI}(\lambda) = \sum_r \log \frac{P_\lambda\left(O_r \middle| W_r\right)}{\sum_{W_k \in \mathbf{W}^r} P_\lambda\left(O_r \middle| W_k\right) P\left(W_k\right)}$$

Training: Maximum Mutual Information (3/4)

- When we maximize the MMIE objection function
  - Not only the probability of true word sequence (numerator, like the MLE objective function) can be increased, but also can the probabilities of other possible word sequences (denominator) be decreased
  - Thus, MMIE attempts to make the correct hypothesis more probable, while at the same time it also attempts to make incorrect hypotheses less probable

Training: Maximum Mutual Information (4/4)

- The objective functions used in discriminative training, such as that of MMI, are often rational functions
  - The original Baum-Welch algorithm is not feasible
  - Gradient descent and the extended Baum-Welch (EB) algorithm are two applicable approaches for such a function optimization problem
    - Gradient descent may require a large number of iterations to obtain an local optimal solution
    - While Baum-Welch algorithm was extended (EB) for the optimization of rational functions
      - MMI training has similar update formulas as those of MPE (Minimum Phone Error) training to be introduced later

Training: Minimum Phone Error (1/4)

- The objective function of Minimum Phone Error (MPE) is also derived with the overall expected risk criterion

$$F_{MPE}(\lambda) = \sum_r \sum_{W_i \in \mathbf{W}^r} p(W_i \mid O_r) A(W_i, W_r)$$

$$\approx \sum_r \sum_{W_i \in \mathbf{W}^r} \frac{p_\lambda(O_r \mid W_i) P(W_i)}{\sum_{W_k \in \mathbf{W}^r} p_\lambda(O_r \mid W_k) P(W_k)} A(W_i, W_r)$$

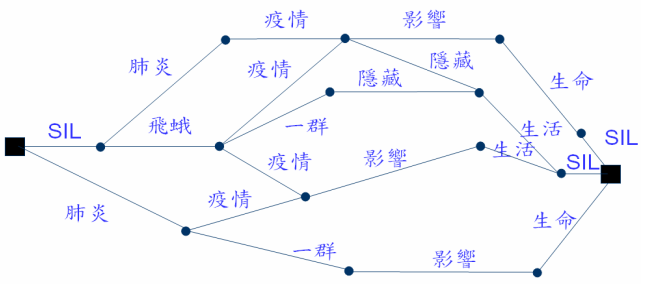
  - Replace the loss function $l(W_i, W_r)$ with the so-called accuracy function $A(W_i, W_r)$ (cf. p.9)

- MPE tries to maximize the expected (phone or word) accuracy of all possible word sequences (generated by the recognizer) regarding the training utterances

Training: Minimum Phone Error (2/4)

- Like the MMI, MPE also aims to make sure that the training data is correctly recognized
  - In MPE each phone error in the training utterance can contribute at most one unit to the objective function, rather than that as an unlimited quantity in MMI (zero-one word string error)
    - The MPE objective function is less sensitive to portions of the training data that are poorly transcribed

$$F_{MMI}(\lambda) = \sum_r \log \frac{P_\lambda(O_r|W_r)}{\sum_{W_k \in \mathbf{W}^r} P_\lambda(O_r|W_k)P(W_k)}$$



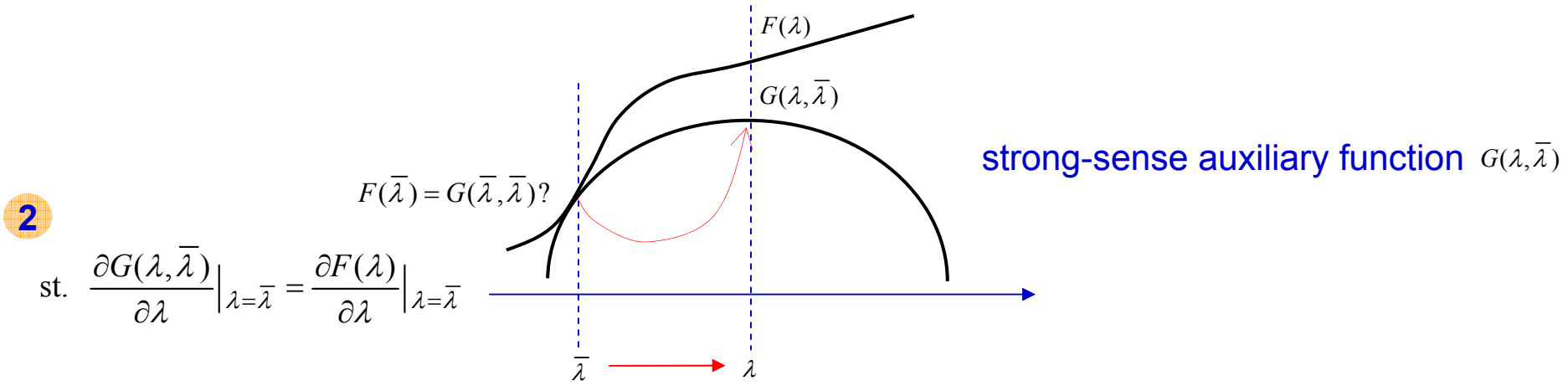
- A (word) lattice structure $\mathbf{W}_{lat}^r$ can be used here to approximate the set $\mathbf{W}^r$ of all possible word sequences of each training utterance
  - Training statistics can be efficiently computed via such a structure

Training: Minimum Phone Error (3/4)

- The objective function $F_{MPE}(\lambda)$ of MPE has the "latent variable" problem, such that it can not be directly optimized
  - However, it is also a rational function and the Baum-Welch (EM) algorithm can not be directly applied
    - Without a strong-sense auxiliary function $G(\lambda,\bar{\lambda})$ satisfying that

1 $$G(\lambda,\bar{\lambda})-G(\bar{\lambda},\bar{\lambda})\le F(\lambda)-F(\bar{\lambda})$$

(or $F(\bar{\lambda})-G(\bar{\lambda},\bar{\lambda})\le F(\lambda)-G(\lambda,\bar{\lambda})$, i,e., $F(\lambda)$and $G(\lambda,\bar{\lambda})$ are the closest when $\lambda=\bar{\lambda}$)



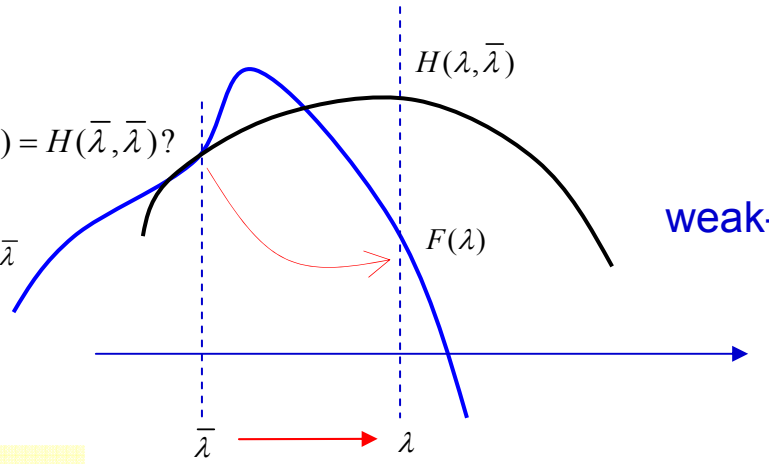
2 st. $$\left.\frac{\partial G(\lambda,\bar{\lambda})}{\partial\lambda}\right|_{\lambda=\bar{\lambda}}=\left.\frac{\partial F(\lambda)}{\partial\lambda}\right|_{\lambda=\bar{\lambda}}$$

Training: Minimum Phone Error (3/4)

- A weak-sense auxiliary function is employed instead for the objective function $F_{MPE}(\lambda)$ of MPE, $H_{MPE}(\lambda,\overline{\lambda})$, which a has a local optimal solution and only satisfies that

1 $$\left.\frac{\partial}{\partial\lambda}H(\lambda,\overline{\lambda})\right|_{\lambda=\overline{\lambda}}=\left.\frac{\partial}{\partial\lambda}F(\lambda)\right|_{\lambda=\overline{\lambda}}$$

st. $\left.\frac{\partial H(\lambda,\overline{\lambda})}{\partial\lambda}\right|_{\lambda=\overline{\lambda}}=\left.\frac{\partial}{\partial\lambda}F(\lambda)\right|_{\lambda=\overline{\lambda}}$



weak-sense auxiliary function $H(\lambda,\overline{\lambda})$

$$F_{MPE}(\lambda)=\sum_r\sum_{W_i\in\mathbf{W}^r}\frac{p_\lambda(O_r\mid W_i)P(W_i)}{\sum_{W_k\in\mathbf{W}^r}p_\lambda(O_r\mid W_k)P(W_k)}A(W_i,W_r)$$

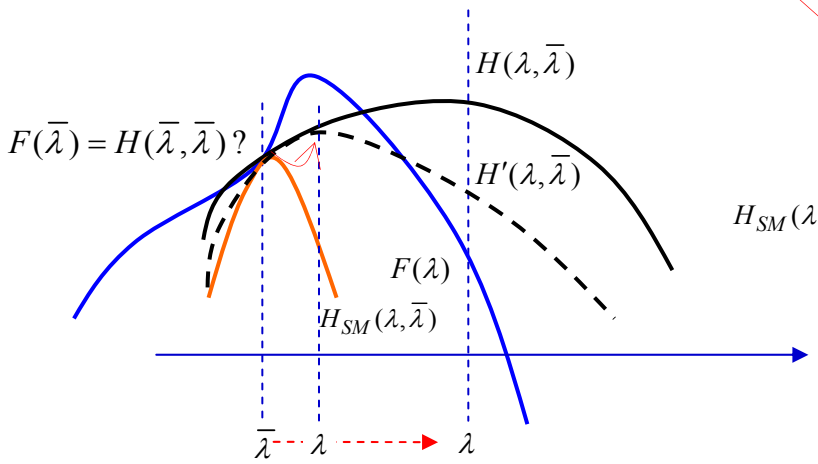
Training: Minimum Phone Error (4/4)

- A smoothing function $H_{SM}(\lambda,\overline{\lambda})$ can be further incorporated into the weak-sense auxiliary function to constrain the function optimization process (to improve convergence)

$$H'_{MPE}(\lambda,\overline{\lambda}) = H_{MPE}(\lambda,\overline{\lambda}) + H_{SM}(\lambda,\overline{\lambda})$$

$$\left[\frac{\partial H_{SM}(\lambda,\overline{\lambda})}{\partial\lambda}\Big|_{\lambda=\overline{\lambda}}\right] = 0$$

st. $\frac{\partial H'(\lambda,\overline{\lambda})}{\partial\lambda}\Big|_{\lambda=\overline{\lambda}} = \frac{\partial}{\partial\lambda}F(\lambda)\Big|_{\lambda=\overline{\lambda}}$



$H_{SM}(\lambda,\overline{\lambda})$ has a optimum value at $\lambda = \overline{\lambda}$

  - The local differential will not be affected and the result ($H'_{MPE}(\lambda,\overline{\lambda})$) is still a weak-sense auxiliary function

MPE: Auxiliary Function (1/2)

- The weak-sense auxiliary function for MPE model updating can be defined as

$$H_{MPE}(\lambda,\overline{\lambda}) = \sum_{r=1}^{R} \sum_{q\in \mathbf{W}^r_{lat}} \left[\frac{\partial F_{MPE}}{\partial \log p_\lambda(O_r \mid q)}\Big|_{\lambda=\overline{\lambda}} \right] \log p(O_r \mid q)$$

$$\frac{\partial F(x)}{\partial x} = \frac{\partial g(x)}{\partial x}\frac{\partial F(x)}{\partial g(x)}$$
$$\frac{\partial H(x)}{\partial x} = \frac{\partial g(x)}{\partial x}\frac{\partial H(x)}{\partial g(x)}$$
$$\frac{\partial (ag(x))}{\partial g(x)} = a$$

  - $\frac{\partial F_{MPE}}{\partial \log p_\lambda(O_r \mid q)}\Big|_{\lambda=\overline{\lambda}}$ is a scalar value (a constant) calculated for each phone arc *q*, and can be either positive or negative (because of the accuracy function)
  - The auxiliary function also can be decomposed as

still have the "latent variable" problem

$$H_{MPE}(\lambda,\overline{\lambda}) = \sum_{r=1}^{R} \sum_{q\in \mathbf{W}^r_{lat}} \max\left(0, \left[\frac{\partial F_{MPE}}{\partial \log p_\lambda(O_r \mid q)}\Big|_{\lambda=\overline{\lambda}} \right]\right) \log p_\lambda(O_r \mid q)$$

arcs with positive contributions (so-called numerator)

$$- \sum_{r=1}^{R} \sum_{q\in \mathbf{W}^r_{lat}} \max\left(0, -\left[\frac{\partial F_{MPE}}{\partial \log p_\lambda(O_r \mid q)}\Big|_{\lambda=\overline{\lambda}} \right]\right) \log p_\lambda(O_r \mid q)$$

arcs with negative contributions (so-called denominator)

$$F_{MPE}(\lambda) = \sum_r \sum_{W_i\in \mathbf{W}_r} \frac{p_\lambda(O_r \mid W_i)P(W_i)}{\sum_{W_k\in \mathbf{W}^r} p_\lambda(O_r \mid W_k)P(W_k)} A(W_i,W_r)$$

These two statistics can be compressed by saving their difference (? I-Smoothing)

MPE: Auxiliary Function (2/2)

- The auxiliary function can be modified by considering the normal auxiliary function $Q_{ML}(\lambda,\overline{\lambda},r,q)$ for $\log p(O_r \mid q)$

$$H_{MPE}(\lambda,\overline{\lambda}) = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \left[\frac{\partial F_{MPE}}{\partial \log p_{\lambda}(O_r \mid q)} \Big|_{\lambda=\overline{\lambda}} \right] Q_{ML}(\lambda,\overline{\lambda},r,q)$$

  - The smoothing term is not added yet here

- The key quantity (statistics value) required in MPE training is $\frac{\partial F_{MPE}}{\partial \log p_{\lambda}(O_r \mid q)} \Big|_{\lambda=\overline{\lambda}}$, which can be termed as

$$\gamma_q^{MPE} = \frac{\partial F_{MPE}}{\partial \log p_{\lambda}(O_r \mid q)} \Big|_{\lambda=\overline{\lambda}}$$

MPE: Statistics Accumulation (1/2)

- The objective function can be expressed as (for a specific phone arc q)

$$F_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_{W_i \in \mathbf{W}_{lat}^r} p_\lambda(O_r \mid W_i) P(W_i) A(W_i, W_r)}{\sum_{W_i \in \mathbf{W}_{lat}^r} p_\lambda(O_r \mid W_k) P(W_k)}$$

$$= \sum_{r=1}^{R} \frac{\sum_{W_i \in \mathbf{W}_{lat}^r, q \in W_i} p_\lambda(O_r \mid W_i) P(W_i) A(W_i, W_r) + \sum_{W_i \in \mathbf{W}_{lat}^r, q \notin W_i} p_\lambda(O_r \mid W_i) P(W_i) A(W_i, W_r)}{\sum_{W_k \in \mathbf{W}_{lat}^r, q \in W_k} p_\lambda(O_r \mid W_i) P(W_i) + \sum_{W_k \in \mathbf{W}_{lat}^r, q \notin W_k} p_\lambda(O_r \mid W_k) P(W_k)}$$

$$\left(= \sum_{r=1}^{R} \frac{a_r(\lambda)}{b_r(\lambda)}\right)$$

- The differential can be expressed as

$$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O \mid q)}\Big|_{\lambda=\bar{\lambda}} = \frac{\partial \frac{a_r(\lambda)}{b_r(\lambda)}}{\partial \log p_\lambda(O \mid q)}\Big|_{\lambda=\bar{\lambda}} = \frac{\frac{\partial a_r(\lambda)}{\partial \log p_\lambda(O \mid q)} b_r(\lambda) - a_r(\lambda) \frac{\partial b_r(\lambda)}{\partial \log p_\lambda(O \mid q)}}{b_r(\lambda)^2}\Big|_{\lambda=\bar{\lambda}}$$

MPE: Statistics Accumulation (2/2)

$$= \left[\frac{\partial a_r(\lambda)}{\partial \log p(O_r \mid q)} \frac{1}{b_r(\lambda)} - \frac{a_r(\lambda)}{b_r(\lambda)} \frac{\partial b_r(\lambda)}{\partial \log p(O_r \mid q)} \frac{1}{b_r(\lambda)} \right]_{\lambda=\bar{\lambda}}$$

$$\left(\because \frac{\partial p_\lambda(O_r \mid W_i)}{\partial \log p_\lambda(O_r \mid q)} = p_\lambda(O_r \mid W_i) \;\; if \;\; q \in W_i \right)$$

$$= \left[\frac{\partial \sum_{W_i \in \mathbf{W}_{lat}^r, q \in W_i} p_\lambda(O_r \mid W_i) P(W_i) A(W_i, W_r)}{\partial \log p_\lambda(O_r \mid q)} \cdot \frac{1}{b_r(\lambda)} - \frac{\partial \sum_{W_i \in \mathbf{W}_{lat}^r, q \in W_k} p_\lambda(O_r \mid W_k) P(W_k)}{\partial \log p(O \mid q)} \cdot \frac{a_r(\lambda)}{b_r(\lambda)^2} \right]_{\lambda=\bar{\lambda}}$$

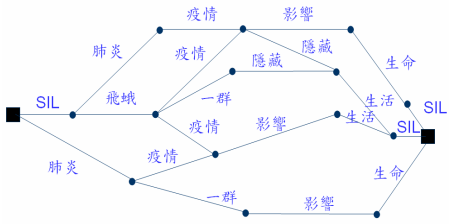
$$K(x) = a \cdot x \cdot c \cdot d$$

$$\frac{\partial K(x)}{\partial \log x} = \frac{\partial x}{\partial \log x} \frac{\partial K(x)}{\partial x} = (x)(a \cdot c \cdot d) = K(x)$$

The average accuracy of sentences passing through the arc q: $c(q)$

γ_q^r: The likelihood of the arc q

$$= \left[\frac{\sum_{W_i \in \mathbf{W}_{lat}^r, q \in W_i} p_{\bar{\lambda}}(O_r \mid W_i) P(W_i) A(W_i, W_r)}{\sum_{W_k \in \mathbf{W}_{lat}^r : q \in W_k} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)} \cdot \frac{\sum_{W_i \in \mathbf{W}_{lat}^r : q \in W_k} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)} - \frac{\sum_{W_i \in \mathbf{W}_{lat}^r, q \in W_k} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)} \cdot \frac{\sum_{W_i \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid W_i) P(W_i) A(W_i, W_r)}{\sum_{W_k \in \mathbf{W}_{lat}^r} p_{\bar{\lambda}}(O_r \mid W_k) P(W_k)} \right]$$



$$= \sum_{r=1}^{R} \gamma_q^r \left(c_r(q) - c_{avg}^r \right)$$

γ_q^r

c_{avg}^r: The average accuracy of all the sentences in the word graph

MPE: Accuracy Function (1/4)

- $c_r(q)$ and c_{avg}^r can be calculated in an approximation way using the word graph and the Forward-Backward algorithm

- Note that the exact accuracy function is express as the sum of phone-level accuracy $A(q)$ over all phones q , e.g.

$$A(q) = \begin{cases} 1 & \text{if correct phone} \\ 0 & \text{if substitution} \\ -1 & \text{if insertion} \end{cases}$$

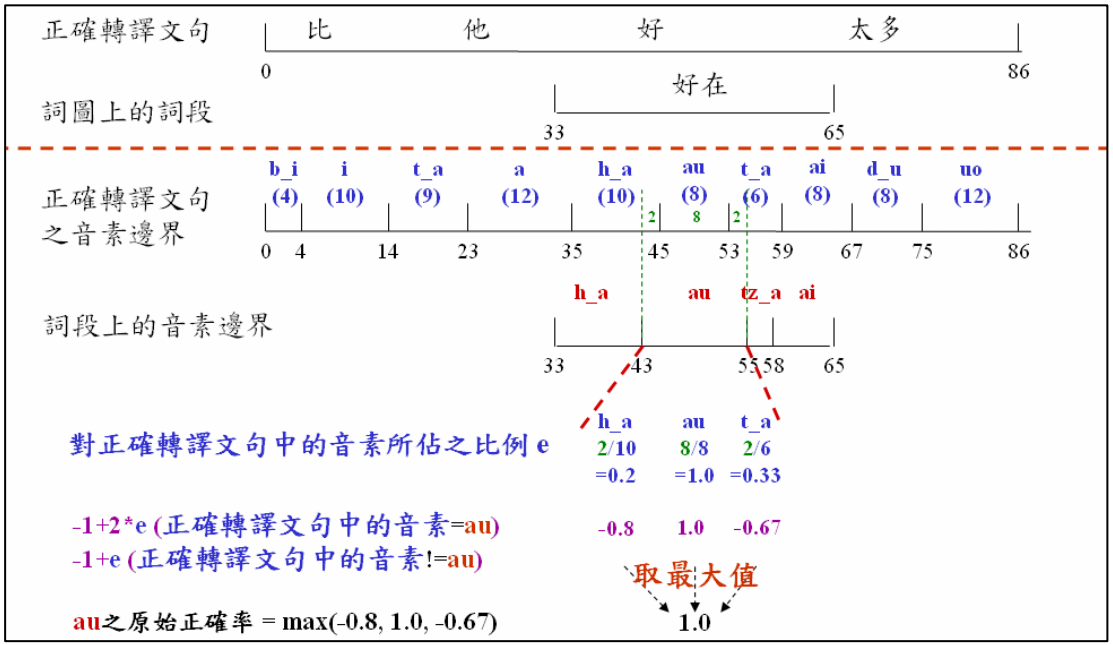
  – However, such accuracy is obtained by full alignment between the true and all possible word sequences, which is computational expensive

MPE: Accuracy Function (2/4)

- An approximated phone accuracy is defined

$$A(q) = \max_{z} \begin{Bmatrix} -1 + 2e(q,z) \text{ if } z \text{ and } q \text{ are same phone} \\ -1 + e(q,z) \quad \text{if different phones} \end{Bmatrix}$$

  - $e(q,z)$: the ration of the portion of z that is overlapped by q



1. Assume the true word sequence has no pronunciation variation
2. Phone accuracy can be obtained by simple local search
3. Context-independent phones can be used for accuracy calculation

MPE: Accuracy Function (3/4)

- Forward-Backward algorithm for statistics calculation
  - Use “phone graph” as the vehicle

Forward

for 開始時間為0的音素q
  $\alpha_q = p(O \mid q)$
  $\alpha'_q = A(q)$
end
for t=1 to T-1
  for 開始時間為t的音素q
    $\alpha_q = 0$
    for 結束時間為t-1且可連至q的音素r
      $\alpha_q = \alpha_q + \alpha_r p(q \mid r)$
    end
    $\alpha'_q = A(q)$
    for 結束時間為t-1且可連至q的音素r
      $\alpha'_q = \alpha'_q + \frac{\alpha_r p(q \mid r)}{\alpha_q} \cdot \alpha'_r$
    end
    $\alpha_q = \alpha_q \cdot p(O \mid q)$
  end
end

MPE: Accuracy Function (4/4)

Backward

for 結束時間為T-1的音素q

$\beta_q = 1$

$\beta'_q = 0$

for t=T-2 to 0

for 結束時間為t的音素q

$\beta_q = 0$

for 開始時間為t+1且可連至q的音素r

$\beta_q = \beta_q + \beta_r p(r|q) p(O|r)$

end

$\beta'_q = 0$

for 開始時間為t+1且可連至q的音素r

$\beta'_q = \beta'_q + \frac{p(r|q) p(O|r) \beta_r}{\beta_q} \cdot (\beta'_r + A(r))$

end

end

end

$$c_{avg} = \frac{\sum_{\text{所有結束時間為 } T-1\text{的音素分枝 } q} \alpha'_q \alpha_q}{\sum_{\text{所有結束時間為 } T-1\text{的音素分枝 } q} \alpha_q}$$

for 每一音素q

$$\gamma_q = \frac{\alpha_q \beta_q}{\sum_q \alpha_q \beta_q}$$

$$c(q) = \alpha'_q + \beta'_q$$

$$\gamma_q^{MPE} = \gamma_q (c(q) - c_{avg})$$

end

MPE: Smoothing Function

- The smoothing function can be defined as

$$H_{SM}(\lambda,\overline{\lambda}) = \sum_{q,m} -\frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|) + (\mu_{qm}-\overline{\mu}_{qm})^T\Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm}\Sigma_{qm}^{-1})\right]$$

  - The old model parameters($\overline{\mu}_{qm}, \overline{\Sigma}_{qm}$) are used here as the parameters
  - It has a maximum value at $\lambda = \overline{\lambda}$

$$\left.\frac{\partial H_{SM}(\lambda,\overline{\lambda})}{\partial \mu_{qm}}\right|_{\substack{\mu_{qm}=\overline{\mu}_{qm}\\ \Sigma_{qm}=\overline{\Sigma}_{qm}}} = -\frac{D_{qm}}{2}\left[2\Sigma_{qm}^{-1}\left(\mu_{qm}-\overline{\mu}_{qm}\right)\right]_{\mu_{qm}=\overline{\mu}_{qm}} = \vec{0}$$

$$\left.\frac{\partial H_{SM}(\lambda,\overline{\lambda})}{\partial \Sigma_{qm}}\right|_{\substack{\mu_{qm}=\overline{\mu}_{qm}\\ \Sigma_{qm}=\overline{\Sigma}_{qm}}} = -\frac{D_{qm}}{2}\left[\Sigma_{qm}^{-T} - \left(\Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm})\right)\left(\Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm})\right)^T - \Sigma_{qm}^{-T}\overline{\Sigma}_{qm}^{T}\Sigma_{qm}^{-T}\right]_{\substack{\mu_{qm}=\overline{\mu}_{qm}\\ \Sigma_{qm}=\overline{\Sigma}_{qm}}} = \vec{0}$$

MPE: Final Auxiliary Function

weak-sense auxiliary function

$$H_{MPE}(\lambda,\overline{\lambda}) = \sum_r \sum_{q\in \mathbf{W}_{lat}^r} \left.\frac{\partial F_{MPE}(\lambda)}{\partial \log\, p_\lambda(O_r \mid q)}\right|_{\lambda=\overline{\lambda}} \log\, p_\lambda(O_r \mid q)$$

strong-sense auxiliary function

$$G_{MPE}(\lambda,\overline{\lambda}) = \sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\, \gamma_{qm}^r(t) \log N(o_r(t), \mu_{qm}, \Sigma_{qm})$$

smoothing function involved

$$G_{MPE}(\lambda,\overline{\lambda}) = \sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\, \gamma_{qm}^r(t) \log N(o_r(t), \mu_{qm}, \Sigma_{qm})$$

weak-sense auxiliary function

$$-\sum_{q,m} \frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|) + (\mu_{qm}-\overline{\mu}_{qm})^T \Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm}\Sigma_{qm}^{-1})\right]$$

MPE: Model Update using EB (1/2)

- Based on the final auxiliary function, the Extended Baum-Welch (EB) training algorithm has the following update formulas

$$\mu_{qm} = \frac{\left\{\theta_{qm}^{num}(O) - \theta_{qm}^{den}(O)\right\} + D_{qm}\bar{\mu}_{qm}}{\left\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\right\} + D_{qm}}$$

$$\Sigma_{qm} = \frac{\left\{\theta_{qm}^{num}(O^2) - \theta_{qm}^{den}(O^2)\right\} + D_{qm}\left[\bar{\Sigma}_{qm} + \bar{\mu}_{qm}\bar{\mu}_{qm}^{T}\right]}{\left\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\right\} + D_{qm}} - \mu_{qm}\mu_{qm}^{T}$$

correlation matrix

q : phone unit
m : mixture component
d : feature dimension

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} o(i)o(i)^{T} - \mu\mu^{T}$$

diagonal covariance matrix

$$\mu_{qmd} = \frac{\{\theta_{qmd}^{num}(O) - \theta_{qmd}^{den}(O)\} + D_{qmd}\bar{\mu}_{qmd}}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}}$$

$$\sigma_{qmd}^{2} = \frac{\{\theta_{qmd}^{num}(O^2) - \theta_{qmd}^{den}(O^2)\} + D_{qmd}(\bar{\sigma}_{qmd}^{2} + \bar{\mu}_{qmd}^{2})}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}} - \mu_{qmd}^{2}$$

MPE: Model Update using EB (2/2)

- Two sets of statistics (numerator, denominator) are accumulated respectively

$$\gamma_{qm}^{num} = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, \gamma_q^{r\,MPE})$$

$$\gamma_{qm}^{den} = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, -\gamma_q^{r\,MPE})$$

$$\theta_{qm}^{num}(O) = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, \gamma_q^{r\,MPE}) O_r(t)$$

$$\theta_{qm}^{den}(O) = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, -\gamma_q^{r\,MPE}) O_r(t)$$

$$\theta_{qm}^{num}(O^2) = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, \gamma_q^{r\,MPE}) O_r(t)^2$$

$$\theta_{qm}^{den}(O^2) = \sum_{r=1}^{R} \sum_{q \in \mathbf{W}_{lat}^{r}} \sum_{t=s_q}^{e_q} \gamma_{qm}^{r}(t) \max(0, -\gamma_q^{r\,MPE}) O_r(t)^2$$

MPE: More About Smoothing Function (1/2)

- The mean and variance update formulas rely on the proper setting of the smoothing constant (D_{qm} or D_{qmd})
  - If D_{qmd} is too large, the step size is small and convergence is slow
  - If D_{qmd} is too small, the algorithm may become unstable
  - D_{qmd} also needs to make all variance σ^2_{qmd} positive

$$\mu_{qmd} = \frac{\theta_{qmd}(O) + D_{qmd}\,\overline{\mu}_{qmd}}{\gamma_{qm} + D_{qmd}} \qquad \sigma^2_{qm} = \frac{\theta_{qmd}(O^2) + D_{qmd}\,[\overline{\sigma}^2_{qmd} + \overline{\mu}^2_{qmd}]}{\gamma_{qm} + D_{qmd}} - \mu^2_{qmd}$$

$$\Rightarrow \frac{\theta_{qmd}(O^2) + D_{qmd}\,[\overline{\sigma}^2_{qmd} + \overline{\mu}^2_{qmd}]}{\gamma_{qm} + D_{qmd}} - \left[\frac{\theta_{qmd}(O) + D_{qmd}\,\overline{\mu}_{qmd}}{\gamma_{qm} + D_{qmd}}\right]^2 > 0$$

$$\Rightarrow (\theta_{qmd}(O^2) + D_{qmd}\,[\overline{\sigma}^2_{qmd} + \overline{\mu}^2_{qmd}])(\gamma_{qm} + D_{qmd}) - (\theta_{qmd}(O) + D_{qmd}\,\overline{\mu}_{qmd})^2 > 0$$

$$\Rightarrow \underbrace{\overline{\sigma}^2_{qmd}\,D^2_{qmd}}_{A} + \underbrace{\left[\theta_{qmd}(O^2) + \overline{\sigma}^2_{qmd}\,\gamma_{qm} + \overline{\mu}^2_{qmd}\,\gamma_{qm} - 2\cdot\theta_{qmd}(O)\overline{\mu}_{qmd}\,d\right]D_{qmd}}_{B} + \underbrace{\theta_{qmd}(O^2)\gamma_{jm} - [\theta_{qmd}(O)]^2}_{C} > 0$$

MPE: More About Smoothing Function (2/2)

$$D_{qmd} > \frac{-B + \sqrt{B^2 - 4 \cdot A \cdot C}}{2 \cdot A}$$

- Previous work [Povey 2004] used a value of D_{qmd} that was twice the minimum positive value needed to insure all variance updates were positive

MPE: I-Smoothing

- I-smoothing increases the weight of the numerator counts depending on the amounts of data available for each Guassian
- This is done by multiplying the numerator terms ($\gamma_{qm}^{num}, \theta_{qm}^{num}(O), \theta_{qm}^{num}(O^2)$) in the update formulas by

$$\theta'^{num}_{qm}(O) = \theta^{num}_{qm}(O) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}} \theta^{ML}_{qm}(O)$$

$$\theta'^{num}_{qm}(O^2) = \theta^{num}_{qm}(O^2) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}} \theta^{ML}_{qm}(O^2)$$

$$\gamma'^{num}_{qm} = \gamma^{num}_{qm} + \tau_{qm}$$

emphasize positive contributions (arcs with higher accuracy)

  - τ_{qm} can be set empirically (e.g., $\tau_{qm} = 100$)

Preliminary Experimental Results (1/4)

- Experiment conducted on the MATBN (TV broadcast news) corpus (field reporters)
  - Training: 34,672 utterances (25.5hrs)
  - Testing: 292 utterances (1.45hrs, outside-testing)
  - Metric: Chinese character error rate (CER)

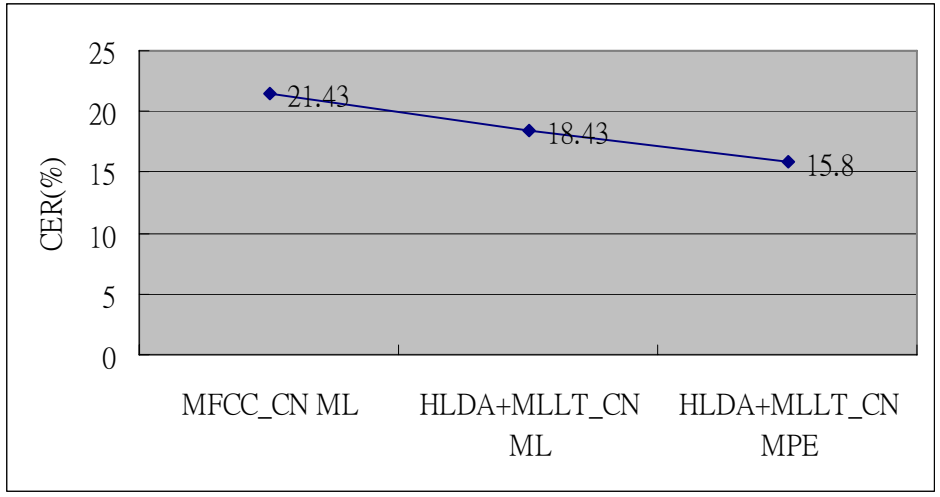
CER(%)	ML_itr10	ML_itr10 MPE_itr10	ML_itr150	ML_itr150 MPE_itr10
MFCC(CMS)	29.48	26.52	28.72	24.44
MFCC(CN)	26.60	25.05	26.96	24.54
HLDA+MLLT(CN)	23.64	**20.92**	23.78	**20.79**

11.51% relative improvement

12.83% relative improvement

Preliminary Experimental Results (2/4)

- Another experiment conducted on the MATBN (TV broadcast news) corpus (field reporters)
  - Discriminative Feature: HLDA+MLLT(CN)
  - Discriminative training: MPE
  - Total 34,672 utterances (24.5hrs) - (10-fold cross validation)



A relative improvement of 26.2% finally achieved (HLDA-MLLT(CN)+ MPE)

Preliminary Experimental Results (3/4)

- Conducted on radio broadcast news recorded from several radio station located at Taipei (anchor speakers)
  - MFCC(CMS)/LDA(CN)/HLDA-MLLT(CN) features
  - Unsupervised Training
  - Metric: Chinese character error rate (CER)

9.94%

	MFCC	MFCC+MPE	LDA	LDA+MPE	HLDA-MLLT	HLDA-MLLT+MPE
Original 4 Hrs	22.56%	-	20.04%	18.13%	20.12%	18.30%
+5 Hrs (Thr=0.9)	18.05%	-	16.43%	14.87%	16.44%	14.78%
+21 Hrs (Thr=0.8)	16.90%	15.34%	15.22%	14.11%	15.64%	14.22%
+33 Hrs (Thr=0.7)	17.32%	-	15.49%	14.23%	15.37%	14.28%
+48 Hrs (Thr=0.6)	17.29%	-	15.54%	14.32%	15.42%	14.31%
+54 Hrs (Thr=0.5)	17.30%	-	15.54%	14.31%	15.46%	14.22%
+60 Hrs (Thr=0.4)	17.17%	-	15.42%	14.22%	15.36%	-

9.32%

16.51%

15.86%

Preliminary Experimental Results (4/4)

- Conducted on microphone speech recorded for spoken dialogues
  - Free syllable decoding with MFCC_CMS features
  - About 8hr training data was used

- Syllable error rate was reduced form 27.79% to 22.74%
  - 18.17% relative improvement

Conclusions

- MPE/MWE (or MMI) based discriminative training approaches have shown effectiveness in Chinese continuous speech recognition
  - Joint training of feature transformation, acoustic models and language models
  - Unsupervised training
  - More in-deep investigation and analysis are needed
  - Exploration of variant accuracy/error functions

References

- D. Povey. “Discriminative Training for Large Vocabulary Speech Recognition,” Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004
- R. Schluter, W. Macherey, B. Muller, H. Ney, “Comparison of discriminative training criteria and optimization methods for speech recognition,” Speech Communication 34, 2001
- K. Vertanen, “An Overview of Discriminative Training for Speech Recognition”
- V. Goel, S. Kumar, W. Byrne, “Segmental minimum Bayes-risk decoding for automatic speech recognition,” IEEE Transactions on Speech and Audio Processing, May 2004
- J.W. Kuo, “An Initial Study on Minimum Phone Error Discriminative Learning of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition” Master Thesis, NTNU, 2005
- J.W. Kuo, B. Chen, "Minimum Word Error Based Discriminative Training of Language Models," Eurospeech 2005

Thank You !