IBIS2009 企画セッション「音響・音声処理と機械学習」

音声系列パターン認識のための識別学習

Discriminative Training for Speech Sequential Pattern Recognition

NTT コミュニケーション科学基礎研究所

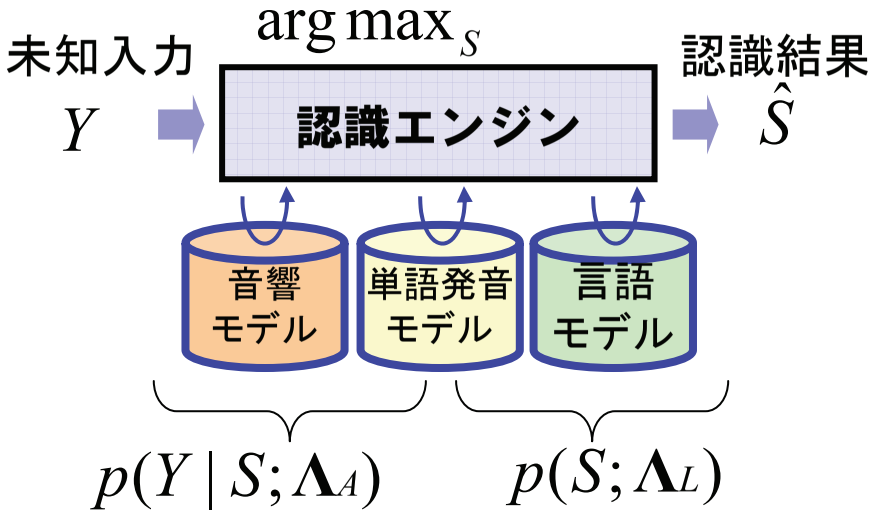
中村 篤

確率的アプローチによる音声認識

- ベイズ決定則に基づく系列探索

$$p(S \mid Y) = \frac{p(Y \mid S)p(S)}{p(Y)}$$

$$\hat{S} = \arg\max_{S} p(S \mid Y)$$
$$= \arg\max_{S} p(Y \mid S)p(S)$$



- 最大尤度基準による(生成)モデルのパラメータ推定 (X: 学習データ)

$$\hat{\Lambda} = \arg\max_{\Lambda} \sum_{X} p(X \mid S_{r,X}; \Lambda_A)\, p(S_{r,X}; \Lambda_L) = \arg\max_{\Lambda_A} \sum_{X} p(X, S_{r,X} \mid \Lambda)$$

$$F_{ML}(X_r, \Lambda) = p(X, S_{r,X} \mid \Lambda) \triangleq p_{\Lambda}(X_r, S_r)$$

X_r : 学習データ特徴量系列

S_r : X_rに対応する正解シンボル系列

「最大尤度」は、識別能力の直接向上につながる基準ではない

発表の内容

音声系列パターン認識において広く用いられている
代表的識別学習手法の概観

最大相互情報量　最小識別誤り　最小シンボル（音素／単語）誤り

成り立ちを異にする各手法について
基本関数（ψー確率）を用いた統一的な解釈・表現

本解釈に基づく各手法の拡張や一般化
ψー確率を中心とする目的関数の関係

識別学習によるモデルパラメータ推定

- 最大相互情報量(MMI: Maximum Mutual Information)基準

[Bahl et al. 86; Valtchev et al. 97]

$$I(X,S)=\sum_{X,S}p(X,S)\log p(S\mid X)-\sum_{X,S}p(X,S)\log p(S)$$

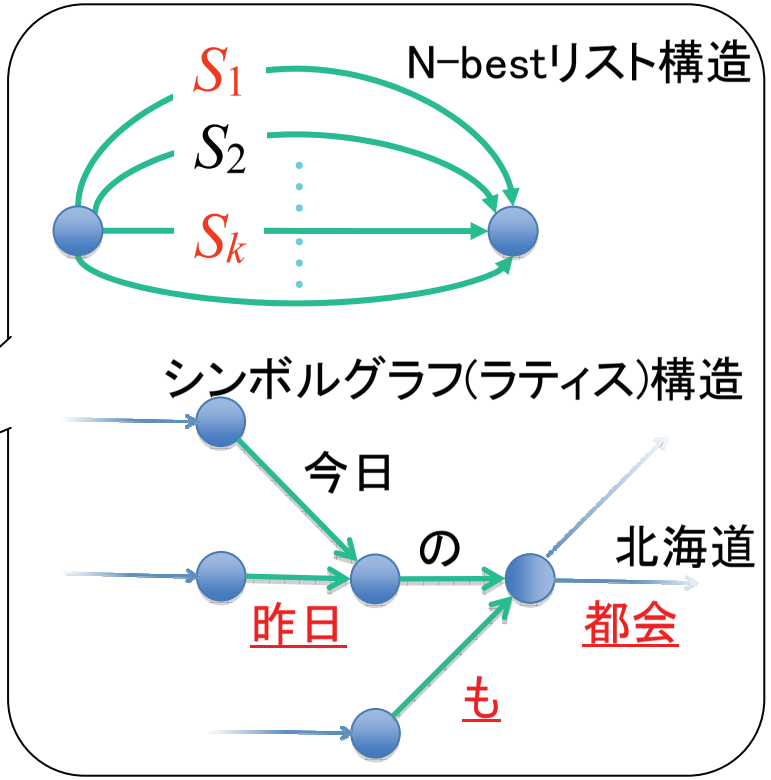
$$=\sum_{X,S}p(X,S)\log\frac{p(X,S)}{p(X)}+H(S)$$

$$\hat{\Lambda}=\arg\max_{\Lambda}\sum_{X}\log\frac{p(X,S_{r,X}\mid\Lambda)}{p(X\mid\Lambda)}$$

$$=\arg\max_{\Lambda}\sum_{X}\log\frac{p(X,S_{r,X}\mid\Lambda)}{\sum_{k}p(X,S_{k}\mid\Lambda)}$$

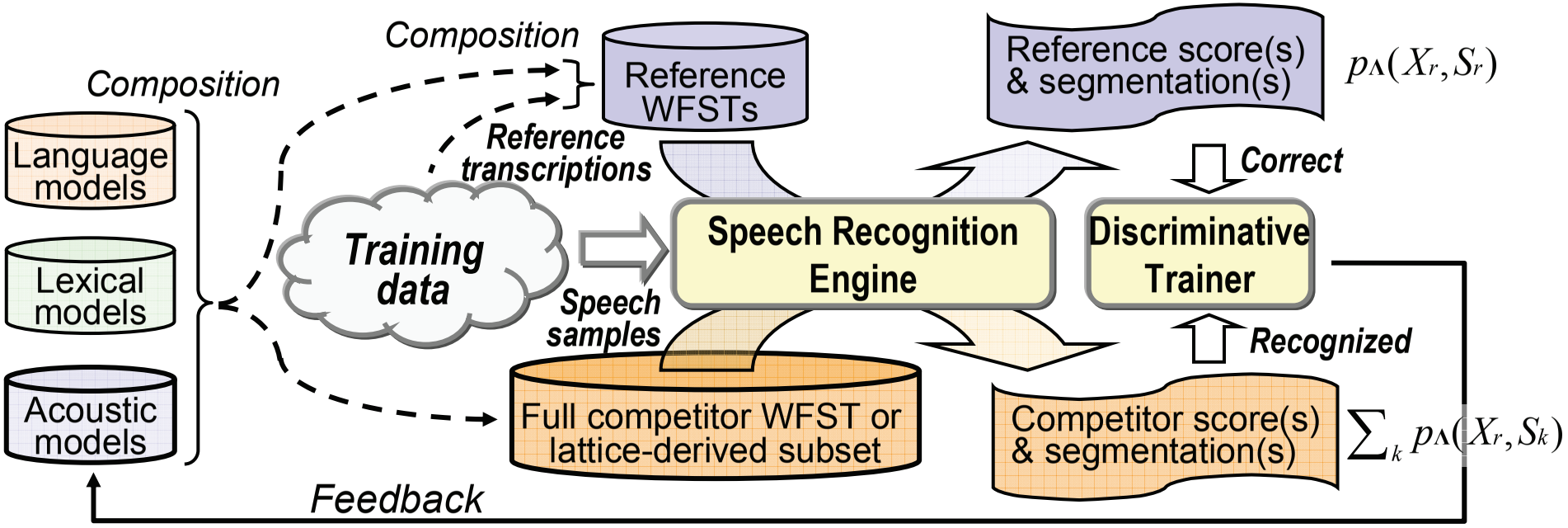
学習データXの認識結果から獲得

$$F_{MMI}(X_r,\Lambda)=\frac{p_{\Lambda}(X_r,S_r)}{\sum_{k}p_{\Lambda}(X_r,S_k)}$$



大語彙識別学習

– 大語彙識別学習の流れ [McDermot et al. 07]



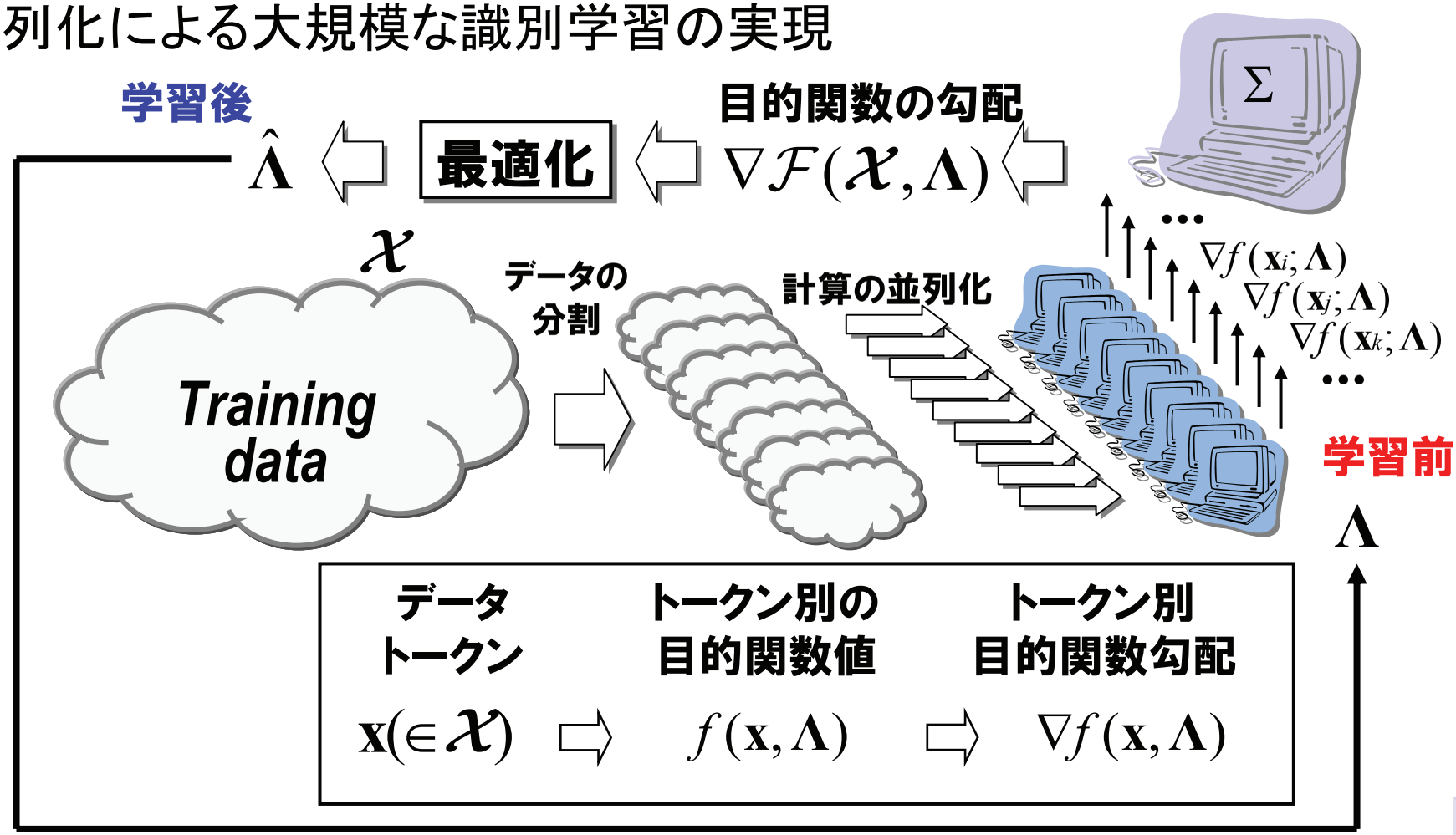
学習データ認識して得られる大規模仮説（N-bestリスト／ラティス）を利用

⇒ 超高速音声認識エンジンの恩恵大

仮説（／ラティスの弧）ごとの偏微分係数（$\partial F_{MMI}/\partial \Lambda_*$）の計算が膨大

大語彙識別学習（つづき）

- 並列化による大規模な識別学習の実現



最小識別誤り学習

- MCE: Minimum Classification Error基準 [Katagiri et al. 92; McDermot et al 97]

識別関数: $\mathcal{G}_i(X_r, \mathbf{\Lambda}) = \log p_{\mathbf{\Lambda}_L}(S_i) + \log p_{\mathbf{\Lambda}_A}(X_r \mid S_i)$

誤分類尺度: $d_r(X_r, \mathbf{\Lambda}) = \underbrace{-\mathcal{G}_r(X_r, \mathbf{\Lambda})}_{\text{正解に対する識別関数値}} + \underbrace{\log\left(\frac{1}{C} \sum_{k \mid S_r \neq S_k} e^{\mathcal{G}_k(X_r, \mathbf{\Lambda}) \cdot \phi} \right)^{\frac{1}{\phi}}}_{\text{不正解に対する識別関数値の一般化平均}}$

損失関数による評価と総和(総損失)の最小化によるパラメータ推定

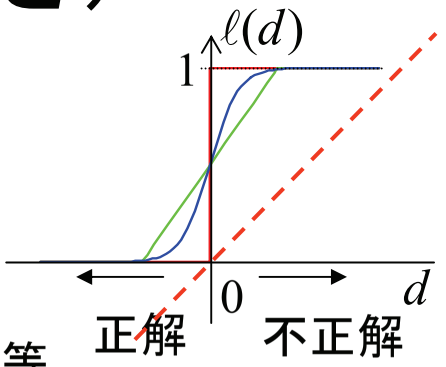
$$\hat{\mathbf{\Lambda}} = \arg\min_{\mathbf{\Lambda}} \sum_{X} \ell(d(X, \mathbf{\Lambda}))$$

正例／負例を陽に区別し、両者の識別関数値の差を元に誤りの度合い（損失）を直接定義・計算

最小識別誤り学習（つづき）

- (シグモイド)損失関数:

$$\ell(d(X,\Lambda)) = \frac{1}{1+\exp(-\eta d(X,\Lambda)+\nu)}$$



※ 他に単位階段／区分線形／ヒンジ／(全域)線形損失関数等

- 誤分類尺度の等価変形

$$d_r(X_r,\Lambda) = \frac{1}{\phi}\left(\log\frac{\sum_{k|Sr\neq Sk} p_{\Lambda L}(S_k)^{\phi}\, p_{\Lambda A}(X_r \mid S_k)^{\phi}}{p_{\Lambda L}(S_r)^{\phi}\, p_{\Lambda A}(X_r \mid S_r)^{\phi}} - \log C\right)$$

$$= \frac{1}{\phi}\left(\log\frac{\sum_{k|Sr\neq Sk} p_{\Lambda}(X_r,S_k)}{p_{\Lambda}(X_r,S_r)} - \log C\right) \triangleq \frac{1}{\phi}\left(\log F_{MCE}(X_r,\Lambda) - \log C\right)$$

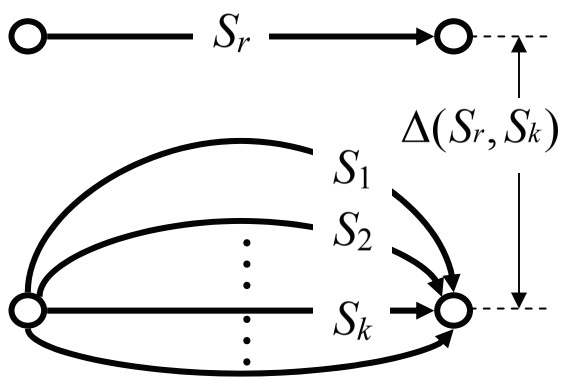
$$F_{MCE}(X_r,\Lambda) = \frac{\sum_{k|Sr\neq Sk} p_{\Lambda}(X_r,S_k)}{p_{\Lambda}(X_r,S_r)}$$

$$\left(\begin{array}{l} p_{\Lambda}(X,S) \\ \triangleq (p_{\Lambda L}(S)\, p_{\Lambda A}(X \mid S))^{\phi} \end{array}\right)$$

Lattice Smoothing factor

最小シンボル(e.g., 音素,単語)誤り学習

- MPE/MWE: Minimum Phone/Word Error [Povey 02]



$$F_{MPE}(X_r, \Lambda) = \frac{\sum_k p_\Lambda(X_r, S_k)\Delta(S_r, S_k)}{\sum_k p_\Lambda(X_r, S_k)} = E[\Delta(S_r, S_k)]$$

$$\frac{\partial F_{MPE}(X_r, \Lambda)}{\partial \log p_\Lambda(X_r, S)} = \frac{p_\Lambda(X_r, S)}{\sum_k p_\Lambda(X_r, S_k)}\left(\underbrace{\Delta(S_r, S) - F_{MPE}(X_r, \Lambda)}\right)$$

$\Delta(\bullet,\bullet)$は相違尺度
正解との相違尺度
＝誤り数

$p_\Lambda(X_r, S)$ が大きくなるよう学習 ← ≤ 0: Sは みなし正解仮説

〃 が小さくなるよう学習 ← > 0: 〃 みなし不正解仮説

「仮説の正/誤」という二値尺度の代りに、「誤り数」という細粒的尺度を導入

陽に正解を意識せず、誤り数の期待値を基準とした大小関係に注目

ラティス上で累積期待値を伝播させる特殊なFW-BWアルゴリズム [Povey 02]

識別的目的関数(主要部)と派生形

- MMI

$$F_{MMI}(X_r, \Lambda) = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)}$$

- MCE

$$F_{MMI}(X_r, \Lambda) = \frac{\sum_{k|\Delta(S_r,S_k)\neq 0} p_\Lambda(X_r, S_k)}{p_\Lambda(X_r, S_r)}$$

- MPE/MWE

$$F_{MPE}(X_r, \Lambda) = \frac{\sum_k p_\Lambda(X_r, S_k)\Delta(S_r, S_k)}{\sum_k p_\Lambda(X_r, S_k)}$$

- Boosted MMI [Povey et al. 08]

$$F_{bMMI}(X_r, \Lambda) = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)\exp(\sigma\Delta(S_r, S_k))}$$

- Boosted MPE/MWE

$$F_{bMPE}(X_r, \Lambda) = \frac{\sum_k p_\Lambda(X_r, S_k)\exp(\sigma\Delta(S_r, S_k))\Delta(S_r, S_k)}{\sum_k p_\Lambda(X_r, S_k)\exp(\sigma\Delta(S_r, S_k))}$$

Boosting factor
仮説のスパースネスを
考慮して確率を補正

識別学習の体系化/相互関係性考察

W. Macherey, et al., “Investigations on error minimizing training criteria for discriminative training in automatic speech recognition,” in *Proc. Eurospeech*, pp. 2133–2136, 2005.

E. McDermott & S. Katagiri: ″Discriminative training via minimization of risk estimates based on Parzen smoothing″, *Journal of Applied Intelligence*, *Kluwer Academic Publishers*, Vol. 25, No. 1, pp. 37–57, August 2006.

G. Heigold, et al., “Modified MMI/MPE: A direct evaluation of the margin in speech recognition,” In *Proc. ICML*, pp. 384–391, 2008.

X. He, et al., “Discriminative learning in sequential pattern recognition,” *IEEE SP Mag.* 25, 5, pp. 14–36, September 2008.

A. Nakamura, et al., ″A unified view for discriminative objective functions based on negative exponential of difference measure between strings, ″ *Proc. ICASSP*, pp. 1633–1636, (2009)

目的関数を統一表現する基本関数

- 文字列間相違尺度の逆指数関数

$$\exp(-\underbrace{\sigma}_{\substack{decay\\ rate}} \cdot \underbrace{\Delta(S_1, S_2)}_{\substack{difference\ measure\\ between\ strings}})$$

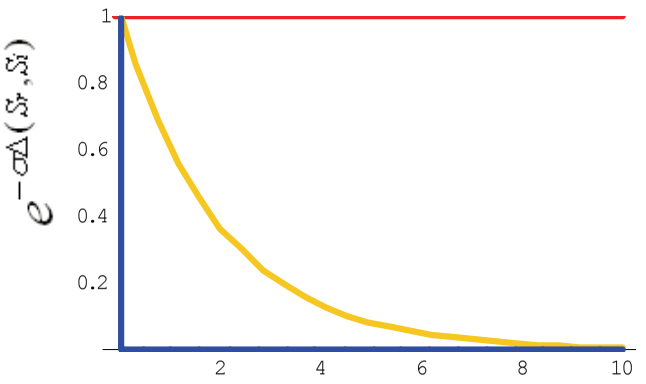
⇒ 指数関数の一般性・解析的取り回しのよさ
細粒的尺度を介した MMI, MCEの定義へ

- 逆指数($\exp(-\sigma\Delta)$)重みつき確率/密度和（ψ-確率）

$$\psi_\sigma(X_r, \Lambda) = \sum_i p_\Lambda(X_r, S_i) \exp(-\sigma\Delta(S_r, S_i))$$

stands for pseudo ($\psi\varepsilon\upsilon\delta o$) probability

$$= \begin{cases} \sum_i p_\Lambda(X_r, S_i) & \sigma = 0 \\ \sum_i p_\Lambda(X_r, S_i) e^{-\sigma\Delta(S_r, S_i)} & \sigma > 0 \\ p_\Lambda(X_r, S_r) & \sigma \to \infty \end{cases}$$



ψ-確率をもとにして、
既存目的関数、それらの拡張・一般化等を広く統一的に表現できる

ψ-確率による既存目的関数の表現 (1)

$$\psi_\sigma(X_r, \mathbf{\Lambda}) = \sum_k p_\Lambda(X_r, S_i) \exp(-\sigma \Delta_k) \qquad \Delta_k = \Delta(S_r, S_k)$$

- MMI

$$\frac{\psi_\infty}{\psi_0} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)} \equiv F_{MMI}(X_r, \mathbf{\Lambda}).$$

- Boosted MMI

$$\frac{\psi_\infty}{\psi_{(-\sigma)}} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k) \exp(\sigma \Delta_k)} \equiv F_{bMMI}(X_r, \mathbf{\Lambda}).$$

- MCE (inverted)

$$\frac{\psi_0}{\psi_\infty - \psi_0} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k) - p_\Lambda(X_r, S_r)} \equiv \frac{1}{F_{MCE}(X_r, \mathbf{\Lambda})}.$$

ψ-確率による既存目的関数の表現 (2)

$$\psi_\sigma(X_r, \mathbf{\Lambda}) = \sum_k p_\Lambda(X_r, S_i) \exp(-\sigma \Delta_k) \qquad \Delta_k = \Delta(S_r, S_k)$$

- MPE/MWE (negated)

$$\frac{\psi'_0}{\psi_0} = -\frac{\sum_k p_\Lambda(X_r, S_k)\Delta_k}{\sum_k p_\Lambda(X_r, S_k)} \equiv -F_{MPE}(X_r, \mathbf{\Lambda})$$

微分操作によって
引っ張り出された
素の誤り数

ただし、 $$\psi'_\sigma = \left.\frac{\partial \psi_\upsilon}{\partial \upsilon}\right|_{\upsilon=\sigma} = -\sum_k p_\Lambda(X_r, S_k) e^{-\sigma\Delta_k} \Delta_k$$

- Boosted MPE/MWE (negated)

$$\frac{\psi'_{(-\sigma)}}{\psi_{(-\sigma)}} = -\frac{\sum_k p_\Lambda(X_r, S_k) \exp(\sigma\Delta_k)\Delta_k}{\sum_k p_\Lambda(X_r, S_k) \exp(\sigma\Delta_k)} \equiv -F_{bMPE}(X_r, \mathbf{\Lambda}).$$

識別的目的関数の拡張・一般化

- 一般化MMI (G-MMI)

$$\frac{\psi_{\sigma 1}}{\psi_{\sigma 2}} = \frac{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma_1 \Delta_k}}{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma_2 \Delta_k}} \qquad (\sigma_1 > \sigma_2)$$

Δに対して急速に減衰（正解の一般化）

Δに対して緩やかに減衰、または増大
（不正解へのBoosting効果）

$\sigma_1 \to \infty, \sigma_2 = 0$ で
(Plain) MMIに一致

$\frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)}$

$p_\Lambda(X_r, S_r)$ ←----- Δ=0の値のみ

$\sum_k p_\Lambda(X_r, S_k)$ ←---- Δの値による減衰無し

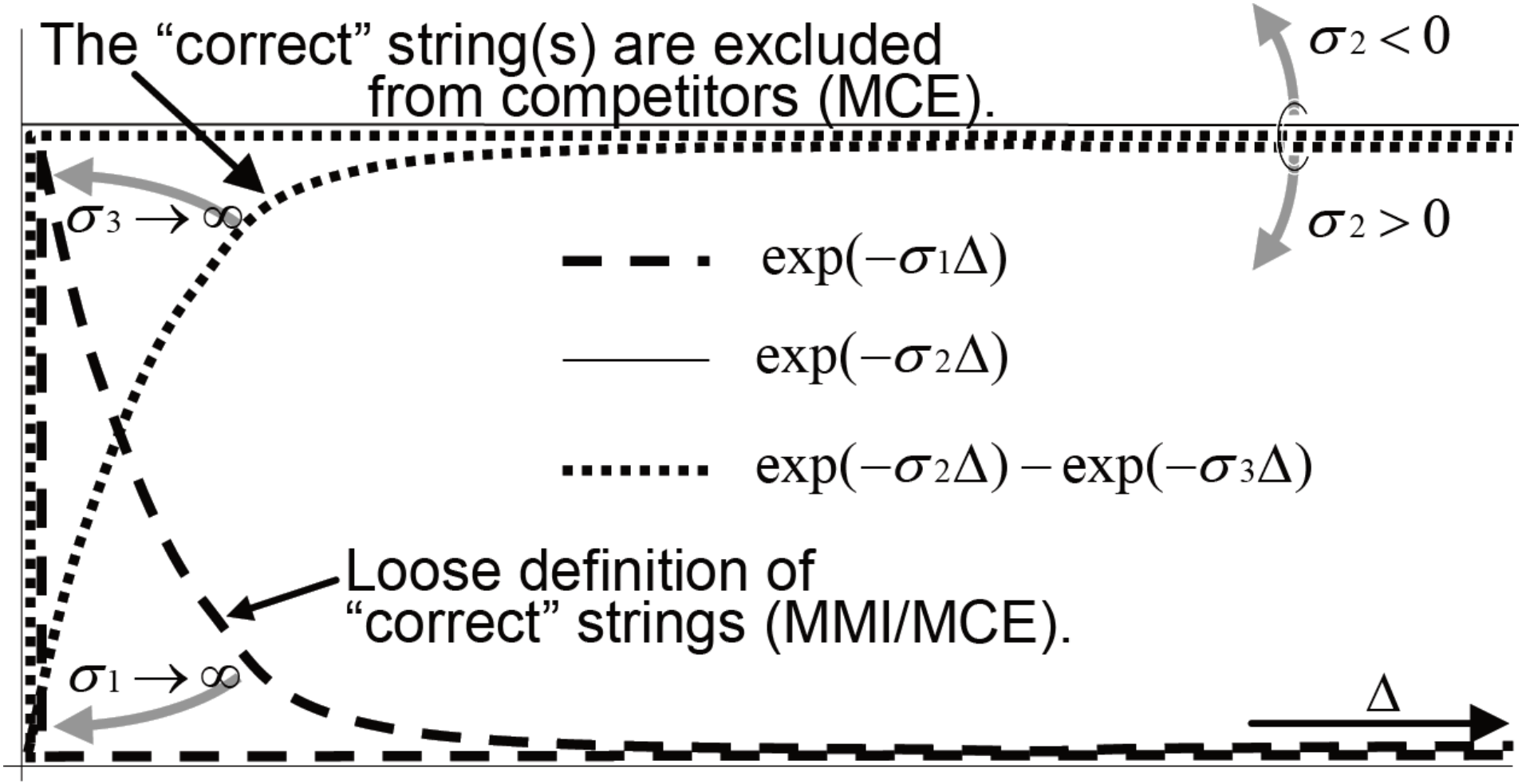
- 一般化MCE (G-MCE; inverted)

$$\frac{\psi_{\sigma 1}}{\psi_{\sigma 2} - \psi_{\sigma 3}} = \frac{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma_1 \Delta_k}}{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma_2 \Delta_k} - \sum_k p_\Lambda(X_r, S_k) e^{-\sigma_3 \Delta_k}}$$

$$(\sigma_1 \simeq \sigma_3 > \sigma_2)$$

$\sigma_1 = \sigma_3 \to \infty, \sigma_2 = 0$ で(Plain) MCEに一致

G-MMI/MCEにおける$\Delta-\psi$減衰特性



最大類似度学習

- 一般化MMI目的関数の等価変形

$$\frac{\psi_{\sigma 1}}{\psi_{\sigma 2}} = \frac{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma 1 \Delta_k}}{\sum_k p_\Lambda(X_r, S_k) e^{-\sigma 2 \Delta_k}} = \frac{\sum_k p_\Lambda(X_r, S_k) e^{\sigma' \Delta_k} e^{-\sigma \Delta_k}}{\sum_k p_\Lambda(X_r, S_k) e^{\sigma' \Delta_k}}$$

$$(\sigma = \sigma_1 - \sigma_2,\ \sigma' = -\sigma_2)$$

- 最大類似度目的関数

$$\sigma' \to 0$$

(0,1]-正規化類似度

$$\frac{\psi_\sigma}{\psi_0} = \frac{\sum_k p_\Lambda(X_r, S_k) \exp(-\sigma \Delta(S_r, S_k))}{\sum_k p_\Lambda(X_r, S_k)} \equiv E[\exp(-\sigma \Delta(S_r, S_k))]$$

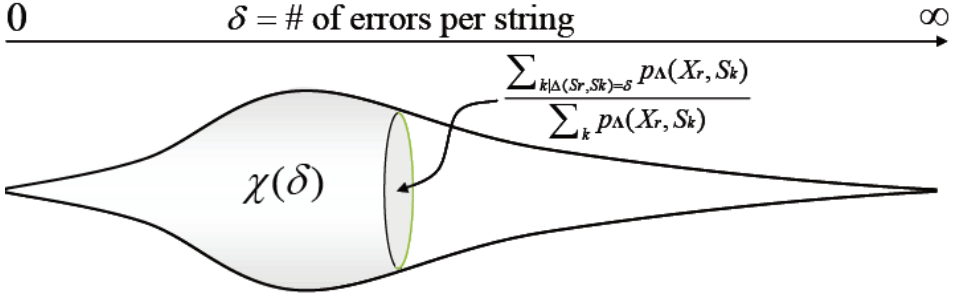
σを適切に（小さく）定めることでMPE/MWEを近似

通常のFW-BWアルゴリズムでラティスに適用可

一般化最小誤りモーメント学習 (1)

- 相違尺度(誤り数)の累積分布

$$\chi(\delta) = \frac{\sum_{k|\Delta(S_r,S_k)\le\delta} p_\Lambda(X_r, S_k)}{\sum_k p_\Lambda(X_r, S_k)}$$



- $\chi(\delta)$のLaplace-Stieltjes変換

$$\underbrace{\mathcal{L}[\chi(\delta)] = \int_0^\infty e^{-\sigma\delta} d\,\chi(\delta)}_{\text{Laplace Stieltjes Transform}} = E\left[\exp(-\sigma\Delta(S_r, S_k))\right] \equiv \underbrace{\frac{\psi_\sigma}{\psi_0}}_{\text{“Maximum Similarity” Objective Function}}$$

$$E\left[\Delta(S_r, S_k)\right] = -\frac{\psi'_0}{\psi_0}, E\left[\left(\Delta(S_r, S_k)\right)^2\right] = \frac{\psi''_0}{\psi_0}, \cdots, E\left[\left(\Delta(S_r, S_k)\right)^n\right] = (-1)^n \frac{\psi^{(n)}_0}{\psi_0}$$

一般化最小誤りモーメント学習 (2)

- 一般化最小誤りモーメント
  (G-MEM: Generalized Minimum Error Moment; negated)基準

$$(-1)^{n-1}\frac{\psi^{(n)}{}_{\sigma}}{\psi_{\sigma}} = \frac{(-1)^{n-1}\left(\frac{\partial}{\partial \upsilon}\right)^{n}\sum_{i} p_{\Lambda}(X_r, S_k)e^{-\upsilon\Delta(S_r,S_k)}\Big|_{\upsilon=\sigma}}{\sum_{k} p_{\Lambda}(X_r, S_k)e^{-\sigma\Delta(S_r,S_k)}}$$

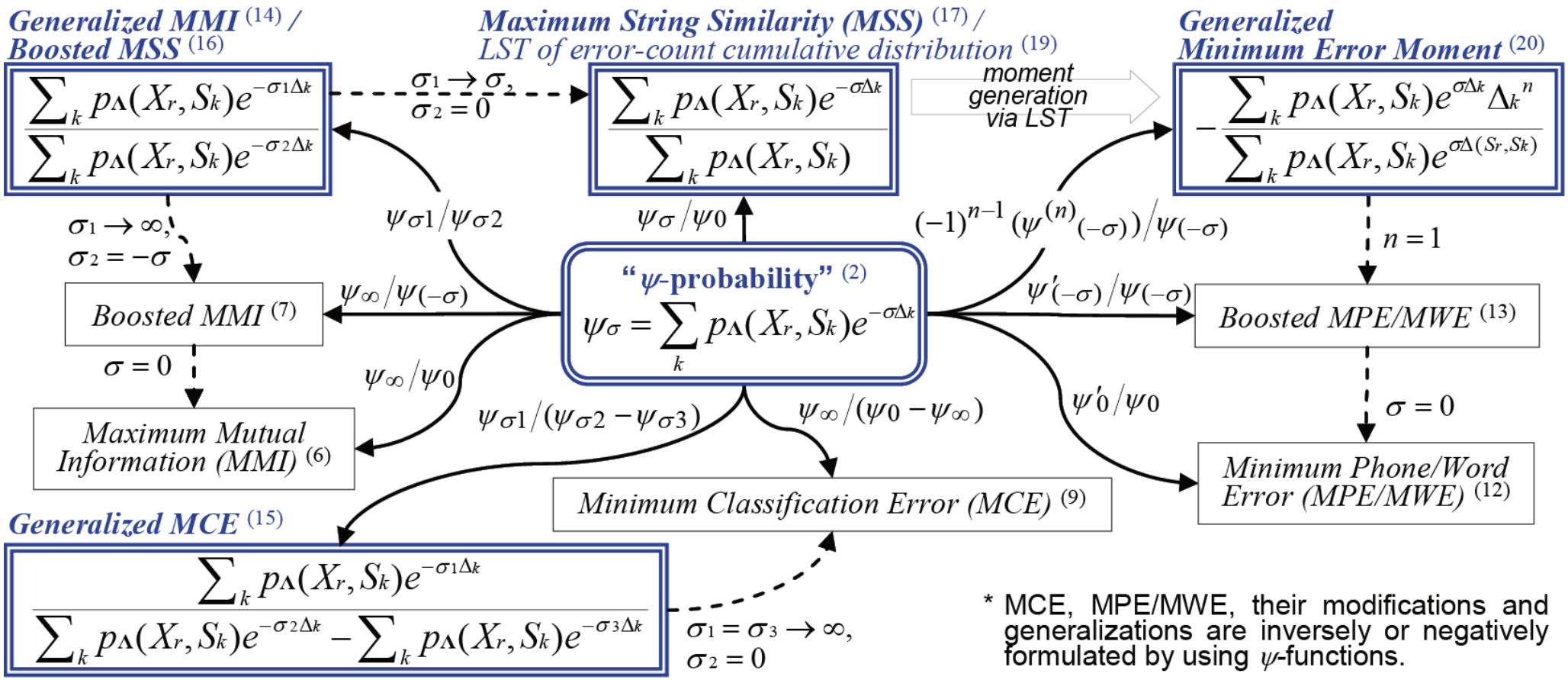
$$= -\frac{\sum_{k} p_{\Lambda}(X_r, S_k)e^{-\sigma\Delta(S_r,S_k)}\left(\Delta(S_r, S_k)\right)^{n}}{\sum_{k} p_{\Lambda}(X_r, S_k)e^{-\sigma\Delta(S_r,S_k)}}$$

$$(n = 1, 2, \cdots;\ \sigma_1 \geq \sigma_2 \geq 0)$$

n=1, σ=0のとき、１次（原点）モーメントとしてMPE/MWE（negated）に一致

誤り数別FW-BWアルゴリズム[McDermott & Nakamura 08]**でラティスに適用可**

識別学習目的関数の関係



まとめ

音声系列パターン認識において広く用いられている、代表的識別学習手法: MMI, MCE, MPE/MWE

成り立ちを異にする各手法について、基本関数(ψ-確率)を用いた統一的な解釈・表現が可能であることを示した

本解釈に基づく各手法の拡張や一般化についても言及し、ψ-確率を中心とする目的関数の関係を示した

以下、補足

ψ-確率: ラティス上で相違尺度(誤り数)が積形式に分解される
⇒ 通常のFW-BWアルゴリズムを適用可

ψ-確率とMPE/MWEの微分解析的関係から数値計算的手法による新たなMPE/MWEの実装を導出 [McDermott et al. 09]