Parameter Adaptation and Compensation in Designing Maximum A Posteriori Decision Rules for Automatic Speech Recognition

Chin-Hui Lee

Multimedia Communication Research Lab
Bell Laboratories, Lucent Technologies
Murray Hill, NJ 07974, USA
chl@research.bell-labs.com

IMA Workshop on Mathematical Foundations for Speech Processing and Recognition

GOALS

- Review advances in adaptive decision rule design based on adaptive decision parameter adaptation for ASR
- Examine what we have learned and why they help ASR work well in many situations
- Give critical view about why ASR does not work as well in many other cases
- Identify areas of real achievement and provide potential research directions to make more advances
OUTLINE

- Statistical Pattern Recognition Paradigm
- Plug-In Decision Rules for Speech Recognition
- Parametric Models and Point Estimation
- Bayesian Approaches to Parameter Adaptation
- On-Line Bayesian Parameter Adaptation
- Structure Parameter Estimation and Adaptation
- Adaptation, Compensation and Robustness
- Robust Decision Rules
- Conclusion

ASR: OPTIMAL BAYES DECISION RULE

- Given $P(X, W)$, the joint distribution of the signal $X$ and the pattern $W$ and a loss function, $\ell(W, d(X))$, of making a decision $d(X)$ when the actual pattern is $W$, then the optimal Bayes decision rule implements

$$d_o(X) = \arg\min_{d(X)} \sum_W \ell(W, d(X)) \cdot P(W|X)$$

- If $\ell(W, d(X))$ is a 0-1 loss function, i.e. error count, then we have the well-known maximum a posteriori decision rule

$$d_{01}(X) = \arg\max_W P(W) \cdot p(X|W)$$

- Difficulties
  - $P(X, W)$ is not known exactly
  - parametric forms of $p_{\Lambda}(X|W)$ and $P_{\Gamma}(W)$ are assumed
  - parameters $\Lambda$ and $\Gamma$ are estimated from training data
ASR: ADAPTIVE DECISION RULES

- Plug-In Maximum A Posteriori (PIMAP) Decoder
  \[ \arg\max_W P(W|X) = \arg\max_W p_A(X|W) \cdot P_T(W) \]

- Key Issues
  - how 'good' are the choices of \( p_A(X|W) \) and \( P_T(W) \)?
  - how 'good' are the point estimators \( \hat{A} \) and \( \hat{T} \)?
  - how 'good' is the PIMAP decoder (any optimality)?

- Bayes Risk Consistency
  - density/parameter estimation consistency implies Bayes risk consistency if the choice of density forms is correct
  - ML/MAP estimates are (large sample) strongly consistent

MESSAGE/SPEECH GENERATION & ASR

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>message source</td>
<td>linguistic channel</td>
<td>articulatory channel</td>
<td>acoustic channel</td>
<td>transmission channel</td>
</tr>
</tbody>
</table>

- message \( M \) realized as a word sequence \( W \)
- words \( W \) realized as a sequence of sound \( S \)
- sounds \( S \) received by transducer in acoustic ambient as \( A \)
- signals \( A \) converted from acoustic to electric, transmitted and received as \( X \) for processing

Sources of Variability

* Task specification * Speaker characteristics
* Contextual variability * Speaking behavior

Speaking environment Transducer variability & distortions
Channel variability and distortions
SOURCE-CHANNEL MODEL FOR ASR

- Simplified ASR: Channel Modeling and Decoding

ESTIMATION OF CLASSIFIER PARAMETERS

- Point estimators to implement the plug-in MAP decoder

- Pro: Hidden Markov Modeling of Speech (and Language)
  - mathematically rigorous, well studied and understood
  - modeling both temporal and spectral variations
  - plenty of textbooks, references, tools (e.g. HTK)
  - data-driven, handling large amounts of training data

- Con: speech is not generated by HMM
  - source of potential robustness problems
  - fallacy of 'there is nothing like more training data'

- Another Perspective: HMM is a discriminant function for performing classification, i.e. computing vs. source model
**HMM ESTIMATION - THREE KEY ADVANCES**

- (1) Detailed Modeling (in many textbooks and references)
  - more data, more context, more mixtures, more tying ...
  - coupled with other techniques, e.g. tree clustering
  - incorporating structures to approximate missing channels

- (2) Adaptive Modeling (this talk)
  - from static to dynamic and on-line classifier design
  - coping with new conditions and unexpected situations

- (3) Discriminative Modeling (next session)
  - from density to decision boundary estimation
  - consistent training and recognition objective

- Many Algorithms - ML, MAP, MDI, MMI, MCE, etc.

---

**ASR CAPABILITIES & LIMITATIONS**

- Use powerful statistical pattern matching paradigms

- Achieve high accuracy if testing data "resemble" what have been seen in training (e.g. TI/NIST CD, RM, ATIS, WSJ)

- Rely on a large application-specific training set to capture all possible speech and language variations (Not Realistic)

- Give high error rate for real-world applications such as in-vehicle hands-free ASR (Not Accurate)

- Imply a degradation in cross-condition testing (Not Robust)

- Reject only a small amount of OOV events (Not Flexible)

- Have a limited understanding capability (Not Intelligent)
GLOBAL AND LOCAL ADAPTATION

Indirect (Global) Adaptation

- Initial models
- Model Transformation (e.g. MLLR)
- Adapted models

Direct (Local) Adaptation

- Initial models
- Model Adaptation
- Adapted models

Joint Adaptation

ADAPTATION APPROACH

- Direct (Local) HMM Adaptation - e.g. Bayesian HMM Adaptation (for VQHMM, TMHMM and CDHMM)
- Indirect (Global) Structure-Based Adaptation
  - MLLR or affine transformation
- Hybrid Structure and HMM Adaptation
  - ML/MAP estimation of both parameter sets
- Structure Correlation and Parameter Tying
- On-Line Incremental Adaptation with Prior Evolution
  - for both parameters and hyperparameters
  - approximate quasi-Bayes approach
BATCH AND ON-LINE ADAPTATION

feature extraction

acoustic normalization

recognizer

adapted models

on-line adaptation

updated hyperparameters

transcription

recognition result

unsupervised

supervised

input speech utterance

ADVANCES IN PARAMETER ADAPTATION

ML-Based Model Transformation/Estimation

- Hierarchical Spectral Clustering (Furu, IEEE ASSP 1995)
- Cepstral Normalization (CDNN) (Acero, 1992)
- Vector Field Smoothing (VFS) (Ohkura et al., ICSLP 92)
- Signal Bias Removal (SBR) (Rahim et al., IEEE SAP 1996)
- Stochastic Matching (SM) (Sankar et al., IEEE SP 1996)
- Non-Linearity SM (Surendran et al., IEEE SAP 1999)

Maximum Likelihood Linear Regression (MLLR) Means only

- Lattner et al., Comp. Speech & Language 1995
- MLLR+Variance Adaptation (Gales et al., Computer Speech & Language 1996)
- Constrained MLLR for HMM Adaptation (Digalakis et al., IEEE SAP 1995)
- MLLR with Dynamic Regression Classes (Lattner et al., ASRPA 2001)

ML-Based Model Transformation/Estimation + MAP-Based Model Adaptation

- MAP-VFS (Tonomura et al., ICASSP 95)
- Incremental MAP-VFS (Takahashi et al., ICASSP 95)
- Model Transformation-MAP (Digalakis et al., IEEE SAP 1996)
- SM-MAP (Chien et al., IEEE SP Letters 1997)
- SM-MAP-VFS (Chien et al., ICASSP 97)

Joint MAP of Transformation and HMM Parameters (Chien et al., ICASSP 2000)

ML Formulation of Model Estimation

- MLE for Mixture HMM (Juang et al., 1996)
- Segmental K-Means Training (Rabiner et al., 1986)
- MLE for Tied Mixture (Huang et al., 1999; Bellegarda, et al., 1999)
- From HMM to Segment Models (Ostendorf et al., 1996)

MAP of CDHMM, Single Gaussian (Lee et al., IEEE SAP 1991)

MAP of CDHMM, Gaussian Mixture F 6 MAP & Segmental MAP (Gaussier et al., 1992, IEEE SAP 1994)

MAP of Correlated Jointly Mean Vectors (Lasey et al., IEEE PAMI 1994)

MAP of Correlated Jointly Mean Vectors (Huang et al., 1989; Bellegarda, et al., 1990)

ML Formulation of Model Estimation

On-Line Adaptive Learning

- MAP of DHMM, SCHMM Empirical Bayes (EB) Learning (Rahim et al., IEEE SAP 1996)
- Adaptive Learning of CDHMM (Huang et al., 1999)
- Prior Evolution (Rahim et al., IEEE PAMI 1994)

MAP of DHMM, SCHMM Empirical Bayes (EB) Learning (Rahim et al., IEEE SAP 1996)

Multiple Linear Regression (Furu, IEEE ASSP 1995)

MAP of Correlated Jointly Mean Vectors (Lasey et al., IEEE PAMI 1994)

Extended MAP (EMAP) Correlation between Gaussian (Zavaliagkos et al., ICASSP 95)

Regression Model Prediction (Cox, CS&L, 1995)

Adaptive Learning, Correlated CDHMM (All Parameters) (Rahim et al., IEEE SP 1996)

Prior Evolution (Rahim et al., Eurospeech 1999)
BAYESIAN HMM ADAPTATION

- Assume HMM parameters random with prior density $p(\Lambda)$
- Investigate three research issues
  - choice of prior density - conjugate prior
  - specification of hyperparameters
  - estimation of parameters - MAP vs. ML estimation
- $\tilde{\Lambda} = \arg\max_{\Lambda} p(\Lambda | X) = \arg\max_{\Lambda} p(X | \Lambda) \cdot p(\Lambda)$
- MLE and MAPE are usually asymptotically equivalent
- Adaptation efficiency and effectiveness
- Batch vs. incremental, supervised vs. unsupervised learning

MAP ESTIMATION EXAMPLES

- Gaussian mean with known variance and prior $\mathcal{N}(\mu, \kappa^2)$
  - $\tilde{\mu} = \frac{TK^2}{\sigma^2 + TK^2} \cdot \overline{x} + \frac{\sigma^2}{\sigma^2 + TK^2} \cdot \mu$
- Gaussian variance with known mean
  - variance clipping to avoid density degeneracy
- Joint Gaussian mean and variance estimation
  - normal-gamma conjugate prior
- Joint multinomial parameter estimation
  - Dirichlet conjugate prior
  - apply to $\pi_i$, $a_{ij}$, $\omega_m$, and other histograms
MAP ESTIMATION OF HMM

- Joint mixture Gaussian estimation of $\theta_k = (\omega_k, m_k, r_k)$
  - product of Dirichlet and normal-Wishart conjugate priors

Let $c_{kt} = \frac{\omega_k N(x_t|m_k, r_k)}{\sum_{l=1}^{K} \omega_l N(x_t|m_l, r_l)}$, then

$$\hat{\omega}_k = \frac{(\nu_k - 1) + \sum_{t=1}^{T} c_{kt}}{\sum_{l=1}^{K} [(\nu_l - 1) + \sum_{t=1}^{T} c_{lt}]}$$

$$\hat{m}_k = \frac{\tau_k \mu_k + \sum_{t=1}^{T} c_{kt} x_t}{\tau_k + \sum_{t=1}^{T} c_{kt}}$$

$$\hat{r}_k = \frac{u_k + \sum_{t=1}^{T} c_{kt} (x_t - m_k)(x_t - m_k)^T + \tau_k (\mu_k - m_k)(\mu_k - m_k)^T}{(\alpha_k - D) + \sum_{t=1}^{T} c_{kt}}$$

- Forward-Backward MAP Estimation of $\lambda_i = (\pi_i, a_{ij}, \theta_{ik})$

- Segmental MAP Estimation -
  $$\hat{\Lambda} = \arg\max_\Lambda \max_s p(X, s|\Lambda) \cdot p(\Lambda)$$

INITIAL PRIOR SPECIFICATION

- Key to the success of Bayesian techniques

- Strict Bayes Approaches
  - known $p(\Lambda|\varphi)$ and the value of $\varphi$ given

- Empirical Bayes Approaches: given $\Lambda_1, \ldots, \Lambda_Q$, estimate $\varphi$
  - method of moment or others
  - prior-weight initialization
  - $\tau$-initialization - from seed models

- More Research Needed

- Important issue of Prior Evolution for On-Line Adaptation
ON-LINE BAYESIAN ADAPTATION

- Recursive Bayes Inference: Non-Reproducible Prior
  \[ p(\Lambda | X^n_1) = \frac{p(X_n | \Lambda) p(\Lambda | X^{n-1}_1)}{\int_{\Omega} p(X_n | \Lambda) p(\Lambda | X^{n-1}_1) d\Lambda} \]

- Quasi-Bayes (QB) Approximation of \( p(\Lambda | X^n_1) \) by \( p(\Lambda | \phi^{(n)}) \)
  \[ p(\Lambda | X^n_1) = \frac{p(X_n | \Lambda) p(\Lambda | \phi^{(n-1)})}{\int_{\Omega} p(X_n | \Lambda) p(\Lambda | \phi^{(n-1)}) d\Lambda} \]

- Prior Evolution based on QB Learning
  - based on recursive evolution of \( p(\Lambda | \phi^{(i)}) \), \( i = 0, ..., L \)
  - approximate posterior with the "most likely" prior
  - incrementally adjust parameters and hyperparameters
  - exponential forgetting and hyperparameter refreshing

- Multiple-Stream Prior Evolution and Posterior Pooling

PRIOR DENSITY APPROXIMATION
ADAPTATION OF STRUCTURE PARAMETERS

- Structure embedded in transformations, e.g. \( \tilde{\Lambda} = F_\phi(\Lambda) \)
  - ML: \( \hat{\phi} = \arg\max_{\phi} p(X|\Lambda, \phi) \)
  - MAP: \( \hat{\phi} = \arg\max_{\phi} p(X|\Lambda, \phi) \cdot p(\phi) \)

- Example: ML/MAP Linear Regression (MLLR/MAPLR)
  - assume \( \hat{m}_k = W_k \cdot m_k \)
  - estimate \( \hat{m}_k \) indirectly through \( W_k \)

- Research Issues
  - how many equivalent class matrices?
  - specification of matrix prior densities
  - unsupervised vs. supervised adaptation
  - on-line adaptation

JOINT PARAMETER ADAPTATION

- ML Estimation of Structure (Nuisance) Parameters followed by MAP Estimation of HMM Parameters

- Joint MAP Estimation of Structure and HMM Parameters
  \( (\tilde{\Lambda}, \hat{\phi}) = \arg\max_{(\Lambda, \phi)} p(X|\Lambda, \phi) \cdot p(\Lambda, \phi) \)

- MAPLR vs. MLLR similar to MAP/HMM vs. ML/HMM

- Joint MAPLR and MAP/HMM better than alone

- Research Issues
  - iterative MAP over \( \Lambda \) and \( \phi \)
  - deal with prior of transformed \( \tilde{\Lambda} = F_\phi(\Lambda) \)
  - distortion introduced by incorrect transformations

- Many new algorithms will follow
STRUCTURE-BASED NORMALIZATION

- Remove irrelevant factors before training/adaptation
- Produce compact speech models
  - cepstral mean normalization (CMN)
  - code dependent cepstral normalization (CDCN)
  - vocal tract length normalization (VTLN)
  - MLLR normalization in speaker adaptive training (SAT)
  - signal bias removal (SBR) and stochastic matching (SM)
- Normalization, Adaptation and Correlation
- Make use of auxiliary structure about missing channels

PARAMETER TYING AND CORRELATION

- Lot of parameters but not enough adaptation data
  - true for both classifier and structure parameters
- Parameter Tying
  - type II and III MAP adaptation
  - tied class adaptation matrices in MLLR
- Parameter Correlation
  - extended MAP (or EMAP)
  - correlated HMM - quasi-Bayes Learning
  - vector field smoothing and MAP/VFS
  - regression based model prediction (RMP)
- Hierarchical Prior Evolution - Structural MAP (SMAP)
MISMATCH BETWEEN TRAINING AND TESTING

\[ \Lambda \]

- \( X \) may be distorted and not easily characterized by \( \Lambda_Y \) causing errors in recognition \( \text{argmax}_W P(W | X, \Lambda_Y) \)
- Form of the distortions \( D_1(\cdot), D_2(\cdot) \) and \( D_3(\cdot) \) may not be known or easily characterized
- Adaptation and Compensation are some solutions

SOURCES OF TRAINING/TESTING MISMATCH

- Microphone and Channel Mismatch
- Changing Channel and Ambient Noise
- Varying Speaker Characteristics and Speaking Style
- Task and Vocabulary Dependency
- Model Incorrectness and Estimation Error
- Combination of Above, Unknown Form of Distortion
ROBUST SPEECH RECOGNITION

- Reduce Training Data Dependency
  - handle testing data not previously seen in training
- Reduce Cross-Condition Mismatch
  - maintain accuracy over a wide range of testing conditions
  - a slight deviation from training conditions should not cause a drastic degradation in ASR performance

COMPENSATION & ADAPTATION

- Fast Adaptation: Make use of adaptation data
  - direct HMM parameter adaptation
  - indirect transformation parameter adaptation
  - MAP widely used for direct adaptation
  - ML widely used for indirect adaptation (e.g. MLLR)
  - combined MAP for both direct and indirect adaptation
- Dynamic Compensation: Make use of testing data
  - auto-adaptation or self-adaptation
  - iterative unsupervised adaptation
  - direct model parameter compensation
- Compensation and adaptation share similar techniques
MAXIMUM-LIKELIHOOD STOCHASTIC MATCHING

- Given trained models \( \Lambda_X \) and a test utterance \( Y \)
- Assume some form of distortion
  
  **Feature Space**: the observed utterance \( Y \) is related to the “original” utterance \( X \) by \( X = F_{\nu}(Y) \)
  
  **Model Space**: the “transformed” model \( \Lambda_Y \) is related to the original models \( \Lambda_X \) by \( \Lambda_Y = G_\eta(\Lambda_X) \)
- Find word string \( W \) and parameters \( \nu \) or \( \eta \) that maximize the joint likelihood \( P(Y, W|\nu \text{ or } \eta, \Lambda_X) \)

ITERATIVE MAXIMIZATION

**Recognition**:

\[
W = \arg\max_W P(W, Y|\nu \text{ or } \eta, \Lambda_X)
\]

**Stochastic Matching**:

**Feature Space Matching**:

\[
\nu = \arg\max_\nu P(Y|\nu, W, \Lambda_X)
\]

**Model Space Matching**:

\[
\eta = \arg\max_\eta P(Y|W, \eta, \Lambda_X)
\]
ROBUST DECISION RULES

- Beyond Plug-In Decision Rules
- Minimax Decision Rules: from Point to Interval Estimate
- Bayesian Predictive Decision Rules: Remove Estimation Uncertainty
- Bayesian Minimax Decision Rules
- Bayesian Predictive Decision Rules Using Structure Parameters
- Other Robust Decision Rules

MINIMAX CLASSIFICATION THEORY

- Partition sample space into decision regions to classify \( X \)
- Assume an uncertainty region \( \Omega_i \) for each HMM \( \Lambda_i \)
- Worst-case probability of error for a decision \( \Omega \)
  \[
  P_\Omega(e) = \sum_{i=1}^{M} p_i \max_{\Lambda \in \Omega_i} \int_{\Omega_i} p_\Lambda(x) dx
  \]
- Minimizing \( P_\Omega(e) \) or its upper bound
  \[
  \hat{P}_\Omega(e) = \sum_{i=1}^{M} p_i \int_{\Omega_i} \max_{\Lambda \in \Omega_i} p_\Lambda(x) dx
  \]
- Equivalently, maximizing
  \[
  1 - \hat{P}_\Omega(e) = \sum_{i=1}^{M} p_i \int_{\Omega_i} \max_{\Lambda \in \Omega_i} p_\Lambda(x) dx
  \]
MINIMAX CLASSIFICATION THEORY (CONT.)

- Two-Step Minimax Classification Solution
  - $\hat{\lambda}_i = \max_{\lambda \in \Lambda_i} p_{\lambda}(x)$
  - $\Omega_i^* = \{x : p_i \cdot \max_{\lambda \in \Lambda_i} p_{\lambda}(x) = \max_j [p_j \cdot \max_{\lambda \in \Lambda_j} p_{\lambda}(x)]\}$
- Feature Space Minimax HMM Inversion is similar

BAYES PREDICTIVE CLASSIFICATION

- Integrating uncertainty in estimating $\Lambda_X$
- Bayes Predictive Classifier (BPC)
  - $\hat{i} = \arg\max_{1 \leq i \leq M} \tilde{p}(C_i|X) = \arg\max_{1 \leq i \leq M} [\tilde{p}(X|C_i) \cdot P(C_i)]$
- Bayes Predictive Density
  - $\tilde{p}(X|C_i) = \int_{\Omega} p(X|\Lambda_i)p(\Lambda_i|\varphi_i)d\Lambda_i$
- Bringing in prior density for $\Lambda_X$: $p(\Lambda_i|\varphi_i)$
- Research Issues
  - quasi-BPC and Viterbi BPC
  - combined on-line adaptation
  - selection of prior density
APPROXIMATE BPC

- Quasi Bayes Predictive Classification (QBPC)
  - normal approximation with \( N(\Lambda_i|\tilde{\Lambda}_i, \tilde{U}_i) \)
  - \( \tilde{p}(X|C_i) \approx p(X|\tilde{\Lambda}_i)p(\tilde{\Lambda}_i|\varphi_i)|\tilde{U}_i|^{1/2} \)
- Viterbi BPC (VBPC)
  - Viterbi segmental approximation
  - \( \tilde{p}(X|C_i) \approx \max_{s,l} p(X, s, l|\tilde{\Lambda}_i)p(\tilde{\Lambda}_i|\varphi_i)|\tilde{U}_i|^{1/2} \)
- On-Line learning of \( \tilde{\Lambda}_i \)
- selection of prior density - normal vs. uniform prior

OTHER ROBUST DECISION RULES

- Approximate Bayes (AB) Rule : embed training data
  \( \hat{W} = \arg\max_W \max_{\Lambda} p(X|\Lambda, W)p(X|\Lambda, W) \frac{P_{\Gamma_0}(W)}{\max_{\Lambda} p(X|\Lambda, W)} \)
- Bayesian Minimax Rule : using MAP instead of ML
  \( \hat{W} = \arg\max_W [p(X|\Lambda_{MAP}, W) \cdot P_{\Gamma_0}(W)] \)
- Using BPC Rule Based on Structure Parameters
  - Bayesian predictive adaptation and compensation
WHY IS ASR HARD?

- Modeling and recognition units are different
- Speech is both nonlinear and nonstationary
  - need simultaneous spectral and temporal modeling
- True models of speech and language unknown
- Interactions between speech and acoustic hard to characterize
- Precise speech distortion models not exactly known
- Sparse training data for speech and language modeling
- Little data to perform adaptation and compensation

SUMMARY

- Plug-In MAP Decision Rules for ASR
- Learning of Classifier Parameters: Most Fruitful Area
  - direct/indirect, ML/MAP adaptive learning
  - on-line incremental and structural learning
- Auxiliary Structure Parameter Estimation
  - improve adaptation efficiency and effectiveness
  - enhance estimation/adaptation through normalization
- Adaptation and Compensation for Robust ASR
- Adaptive Robust Decision Rules
- Knowing interactions amongst speech, language and acoustics is key, no single solution will solve all the problems