A decision Theoretic Formulation for Robust Automatic Speech Recognition (2)

Author : Qiang Huo

Reporter : CHEN TZAN HWEI
The goal of speech recognition can be viewed as a decision problem. i.e. based on the information of $X$, we attempted to make the best decision of the word sequence $W$ that has been embedded in $X$.

For the simplicity of discussion, we can view each $W$ as a class. So, speech recognition consists to find optimal decision rules for classification of the observation $X$ into one of some fixed classes.
In this framework, the issue of constructing an optimal decision rule becomes the following loss minimization problem:

\[
\min_{d(\cdot) \in D} r(d(\cdot)) = \min_{d(\cdot) \in D} \int p(X) \left[ \sum_{W \in \Omega_W} \ell(W, d(X)) p(W | X) \right] dX \quad (5)
\]

The optimization can be solved by minimizing the expression in the square brackets:

\[
d_o(X) = \arg \min_{d(X) \in \Omega_W} \sum_{W \in \Omega_W} \ell(W, d(X)) p(W | X) \quad (6) \text{ Bayes' decision rule}
\]
In summary, in constructing these optimal decision rules, it was assumed that complete prior information about the classes is known:

- The observation space $\Omega_x$ is given
- The loss function $\ell(W, d(X))$ is given
- The true PDF $p(W, X)$ or $p(X \mid W)$ and $p(W)$ are given
Violations of modeling assumption in ASR (1/2)

- Three main distortion types
  - Distortion caused by small-sample effects
  - Distortion of models or discriminant functions for training samples
  - Distortion of trained model or discriminant functions for observation to be classified.
Violations of modeling assumption in ASR (2/2)

- Toward adaptive and robust ASR:
  - Find invariant features so as to minimize the observation variability.
  - Adapting recognizer parameters to new operating conditions using adaptation and/or testing data.
  - Using robust decision strategies
  - Possible combinations of the above techniques.
Adapting recognizer parameters (1/2)

- There must exist a true distribution $p(W, X)$:
  - A solution to improving the adaptive decision rules is to collect training data $\mathcal{X}_a = \{W_a^i, X_a^i\}; i = 1, 2, \ldots, N_a$

Problem: to deal with the problem of estimating a large number of parameters

- Regularization
- Imposing constraints
Adapting recognizer parameters (2/2)

- Regularization:
  - Ex: MAP adaptation

\[
\hat{\Lambda}_{ML} = \max_{\Lambda} p(X | \Lambda) \rightarrow \text{Maximum Likelihood Estimation}
\]

\[
\hat{\Lambda}_{MAP} = \max_{\Lambda} p(\Lambda | X) = \max_{\Lambda} \frac{p(X | \Lambda)p(\Lambda)}{p(X)} = \max_{\Lambda} p(X | \Lambda)p(\Lambda)
\]

- Imposing constraint:
  - Ex: transformed-based (MLLR)

\[
\Lambda_{TB} = F_{\Phi}(\Lambda)
\]
Robust decision rules (1/10)

- The classification performance of the decision rule in a situation are fitted to the distorted model $M_\varepsilon \in M_\varepsilon^*$ is

$$r_\varepsilon (d(\cdot)) = E[\ell(W, d(X))]$$

- Let's define two functional risk:
  - Guaranteed (upper) risk: $r_+ (d(\cdot)) = \sup_{M_\varepsilon \in M_\varepsilon^*} r_\varepsilon (d(\cdot))$
  - Overall risk: $\tilde{r}(d(\cdot)) = E[r_\varepsilon (d(\cdot))]$
Robust decision rules (2/10)

- There are two optimality criteria in searching robust decision rules
  - Minimax decision rule: $d_+(\cdot) = \arg\min_{d(\cdot)} r_+(d(\cdot))$
  - Predictive decision rule: $\tilde{d}(\cdot) = \arg\min_{d(\cdot)} \tilde{r}(d(\cdot))$
Robust decision rules (3/10)

- Both of them assume that
  - The distribution $p(X | W)$ and $p(W)$ are known up to some specifiable parameters in the form of $p_\Lambda(X | W)$ and $p_\Gamma(W)$
  - The true parameters of these distributions, $\Lambda$ and $\Gamma$ lie in a neighborhood of the estimated ones.
Robust decision rules (4/10)

- Minimax decision rules

Let $\eta_\varepsilon(\Lambda_0, \Gamma_0)$ denote the uncertainty neighborhood of the true model parameters $\Lambda, \Gamma$, i.e.,

$(\Lambda, \Gamma) \in \eta_\varepsilon(\Lambda_0, \Gamma_0)$.

Then, we have

$M^*_\varepsilon = \{ p_\Lambda(X \mid W), p_\Gamma(W) \mid (\Lambda, \Gamma) \in \eta_\varepsilon(\Lambda_0, \Gamma_0) \}$

$$r_+ (d(\cdot)) = \sup_{(\Lambda, \Gamma) \in \eta_\varepsilon(\Lambda_0, \Gamma_0)} \sum_{W \in \Omega_W} p_\Gamma(W) \int_{X \in \Omega_X} \ell(W, d(X)) p_\Lambda(X \mid W) dX$$

$$r_{++} (d(\cdot)) = \sum_{W \in \Omega_W} p_{\Gamma_0}(W) \int_{X \notin \Omega_X(W)} \sup_{(\Lambda) \in \eta_\varepsilon(\Lambda_0)} p_\Lambda(X \mid W) dX$$

$$d_{++}(X) = \arg \max_W \left[ p_{\Gamma_0}(W) \sup_{(\Lambda) \in \eta_\varepsilon(\Lambda_0)} p_\Lambda(X \mid W) \right]$$

2006/3/7      NTNU SPEECH LAB  12
Robust decision rules (5/10)

- Minimax decision rules (cont)

  It can be solved in two steps

To estimate the underlying parameters using the ML approach within each neighborhood $\eta_e(A_0^{(W)})$

$$\Lambda_w = \arg\max_{\Lambda_{\eta_e(A_0^{(W)})}} (p_{\Lambda} (X | W))$$

Then, we apply the Plug - in MAP decision rule with $\Lambda_w$ replacing the original $A_0^{(W)}$
Robust decision rules (6/10)

- Predictive decision rule
  - Our prior knowledge about $(\Lambda, \Gamma)$ is assumed a general prior PDF $p(\Lambda, \Gamma | \psi_\Lambda^0, \psi_\Gamma^0)$.
  - Further assume $p(\Lambda, \Gamma | \psi_\Lambda^0, \psi_\Gamma^0) = p(\Lambda | \psi_\Lambda^0) \cdot p(\Gamma | \psi_\Gamma^0)$
  - Often referred as a Bayesian predictive classification rule.
Robust decision rules (7/10)

- Predictive decision rule
- There are some way to evolve $p(\Lambda, \Gamma)$
  - Given a training set $\chi$
    $$p(\Lambda, \Gamma \mid \chi) = \frac{p(\chi \mid \Lambda, \Gamma)p(\Lambda, \Gamma \mid \psi^0_{\Lambda}, \psi^0_{\Gamma})}{\int_{\Omega_{\Lambda}} \int_{\Omega_{\Gamma}} p(\chi \mid \Lambda, \Gamma)p(\Lambda, \Gamma \mid \psi^0_{\Lambda}, \psi^0_{\Gamma})d\Lambda d\Gamma}$$
    $$= p(\Lambda \mid \chi)p(\Gamma \mid \chi)$$
  - A more flexible empirical Bayes approach in which a specific parametric PDF
    $$p(\Lambda, \Gamma \mid \psi_{\Lambda}, \psi_{\Gamma}) = p(\Lambda \mid \psi_{\Lambda})p(\Gamma \mid \psi_{\Gamma})$$
Robust decision rules (8/10)

Predictive decision rule (cont)

A more flexible empirical Bayes approach in which a specific parametric PDF (cont)

We consider the distorted set of model $M^*_\varepsilon$:

$$M^*_\varepsilon = \{p_\Lambda(X | W), p_\Gamma(W) | (\Lambda, \Gamma) \sim p(\Lambda, \Gamma | \psi_\Lambda, \psi_\Gamma); \Lambda \in \Omega_\Lambda, \Gamma \in \Omega_\Gamma\}$$

Based on the above $M^*_\varepsilon$

$$r(d(\cdot)) = E_{(w,X)}E_{(\Lambda,\Gamma)}[\ell(W, d(X))]$$

$$= \sum_{W \in \Omega_W} \int_{X \in \Omega_X} \int_{\Lambda \in \Omega_\Lambda} \int_{\Gamma \in \Omega_\Gamma} \ell(W, d(X))p(W, X | \Lambda, \Gamma)p(\Lambda, \Gamma | \psi_\Lambda, \psi_\Gamma)d\Gamma d\Lambda dX$$
Robust decision rules (9/10)

- Predictive decision rule (cont)
  - A more flexible empirical Bayes approach in which a specific parametric PDF (cont)

\[
\tilde{p}(X | W) = \int_{\Lambda \in \Omega_{\Lambda}} p(X | \Lambda, W) p(\Lambda | \psi_{\Lambda}) d\Lambda
\]

\[
\tilde{p}(W) = \int_{\Gamma \in \Omega_{\Gamma}} p(W | \Gamma) p(\Gamma | \psi_{\Gamma}) d\Gamma
\]

are called predictive densities.

Then, under the \((0,1)\) - loss function, the predictive decision rule

\[
\tilde{d}(X) = \arg \max_{W} \tilde{p}(X | W) \tilde{p}(W)
\]

is referred to as the Bayesian predictive classification (BPC) rule.
Robust decision rules (10/10)

- Three key issues arise in BPC:
  - The definition of the prior density \( \rho(\Lambda, \Gamma | \psi_\Lambda, \psi_\Gamma) \) for modeling the uncertainty of the model parameters \( \Lambda \) and \( \Gamma \)
  - The specification of the hyperparameters \( \psi_\Lambda \) and \( \psi_\Gamma \)
  - The evaluation of the predictive density
Summary

- In this chapter, we have explained several key concepts about
  - The optimal decision rule
  - Adaptive decision rule
  - Robust decision rule

- All of the decision rules described in the chapter aim at achieving the minimum classification error of $W$ instead of the WER.