論文 / 著書情報
Article / Book Information

Title	Predictive Minimum Bayes Risk Classification for Robust Speech Recognition
Authors	Jen-Tzung Chien, Koichi Shinoda, SADAOKI FURUI
Citation	Proc. INTERSPEECH2007, Vol. , No. , pp. 1062-1065,
Pub. date	2007, 8
Copyright	(c) 2007 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Predictive Minimum Bayes Risk Classification for Robust Speech Recognition

Jen-Tzung Chien[a], *Koichi Shinoda*[b] *and Sadaoki Furui*[b]

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
[b] Department of Computer Science, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, Japan
jtchien@mail.ncku.edu.tw, {shinoda,furui}@cs.titech.ac.jp

Abstract

This paper presents a new Bayes classification rule towards minimizing the *predictive Bayes risk* for robust speech recognition. Conventionally, the plug-in maximum *a posteriori* (MAP) classification is constructed by adopting nonparametric loss function and deterministic model parameters. Speech recognition performance is limited due to the environmental mismatch and the ill-posed model. Concerning these issues, we develop the *predictive minimum Bayes risk* (PMBR) classification where the predictive distributions are inherent in Bayes risk. More specifically, we exploit the *Bayes loss function* and the *predictive word posterior probability* for Bayes classification. Model mismatch and randomness are compensated to improve generalization capability in speech recognition. In the experiments on car speech recognition, we estimate the prior densities of hidden Markov model parameters from adaptation data. With the prior knowledge of new environment and model uncertainty, PMBR classification is realized and evaluated to be better than MAP, MBR and Bayesian predictive classification.

Index Terms: Bayes classification, predictive distribution, robust speech recognition

1. Introduction

Automatic speech recognition plays a crucial role in systems that let people communicate naturally with machines, and many investigators have contributed to the resolution of different issues. Among these issues, how to build a robust decision algorithm is critical. Maximum likelihood (ML) training combined with plug-in maximum *a posteriori* (MAP) decision has been used in most applications, but this approach has not met application-specific requirements or proved robust to environmental variations. It is necessary to develop adaptive algorithms for different requirements and variations. Goel et al. [5] have developed the minimum Bayes risk (MBR) classification through introducing word error rate (WER) loss function. MBR decoding outperformed MAP decoding with equal loss for different misclassifications. Minimax classification [8] and Bayesian predictive classification (BPC) [1][6] have also been used to make decision rules robust to environmental variations. Minimax decision rules can assure the smallest maximum risk for all admissible variations, while BPC decision rules guarantees the smallest overall risk.

In this paper, we survey a series of decision rules and present a new Bayes classification rule by minimizing the expected loss where a *loss function* and a *word posterior probability* are embedded. Basically, loss function acts as a classification penalty. In Goel's MBR decoding [5], a WER loss function was determined by matching the assumed target string and the hypothesized string by dynamic programming scheme. The lattice-based word error minimization was implemented [9]. Also, the word posterior probability was viewed as a confidence measure in decoding algorithm [10]. The best segmentation and the segmentations from other competing candidates were merged in calculation of posterior probability [3].

In previous studies, the loss function and the word posterior probability were determined disregarding the issues of adverse condition and ill-posed model. To conduct a Bayesian treatment, we characterize the uncertainty of speech hidden Markov model (HMM) using prior distribution and merge it in calculation of Bayes risk. We present the Bayes loss function and the predictive word posterior probability for MBR decoding using predictive distributions. Loss due to a misclassification is viewed as a hypothesis-test problem. Through examining the hypotheses of loss and lossless events, the Bayes loss [2] is established by a function of *predictive distribution ratio*. These prior distributions are also used in calculation of predictive word posterior probability. We follow the spirit of Bayes theory and establish the *predictive minimum Bayes risk* (PMBR) classification for robust speech recognition. The prior uncertainty is estimated from adaptation data and associated with the specific speaker and noise condition. PMBR classification is built for prediction of unknown test data. Experiments on noisy speech recognition were used to evaluate the effects of prior density estimation and predictive distribution in the performance of decoding algorithms.

2. Bayes Classification Rules

Because of great success of HMMs and *n*-gram models, a variety of statistical approaches have been developed for speech recognition. Conventionally, ML estimation was used to train HMM $\hat{\Lambda}$ and *n*-gram $\hat{\Gamma}$ as *point estimates*. Trained parameters were plugged in MAP decision to transcribe test sentence X into the word sequence $\hat{W}$

$$d_{\mathrm{MAP}}(X) = \hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} P_{\hat{\Lambda}}(X|W) P_{\hat{\Gamma}}(W) . \quad (1)$$

However, a full Bayes decision should be achieved by minimizing the overall risk, or the expectation of loss function $l(W, d(X))$ with respect to $W \in \Omega_W$ and $X \in \Omega_X$

$$\begin{aligned} E_{(W,X)}[l(W,d(X))] &= \int_{X\in\Omega_X} P(X) \left[\sum_{W\in\Omega_W} l(W,d(X)) P(W|X) \right] dX \\ &= \sum_{W\in\Omega_W} P_{\Gamma}(W) \int_{X\in\Omega_X} l(W,d(X)) P_{\Lambda}(X|W) dX \equiv r(d(\cdot)) . \end{aligned} \quad (2)$$

MBR decision is constructed by

$$\begin{aligned} d_{\mathrm{MBR}}(X) &= \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} l(W,d(X)) P(W|X) \\ &= \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} P_{\Gamma}(W) \int_{X\in\Omega_X} l(W,d(X)) P_{\Lambda}(X|W) dX . \end{aligned} \quad (3)$$

Because several assumptions are made in implementation of MBR decision, speech recognition performance is limited substantially.

One is that the observation space Ω_X is known, another is that the distributions of the acoustic model $P_\Lambda(X|W)$ and language model $P_\Gamma(W)$ are known, and the third is that the loss function $l(W,d(X))$ is not considered. MAP decision in (1) is obtained without considering loss for different misclassification, or adopting a zero-one loss function $l_{01}(W,d(X))$. In this study, we are engaged in handling three assumptions for robust speech recognition. The first assumption is problematic when the training data does not match the test data. An adaptive decision is made via adapting hyperparameters to the unknown observation space [4]. For the remaining two assumptions, we introduce some classification rules.

2.1. Minimax and BPC Classification

With regard to the second assumption, when the assumed distributions $P(X|W)$ and $P(W)$ are not consistent with the true ones, the parameters Λ and Γ shall incur estimation errors. We can compensate these errors by using minimax [8] and Bayesian predictive classification (BPC) [1][6] decision rules. Let $\eta(\hat{\Lambda},\hat{\Gamma})$ denote the uncertainty region of the true parameters Λ,Γ where $\hat{\Lambda},\hat{\Gamma}$ are ML parameters. A minimax decision rule is intended to minimize the guaranteed upper risk in the uncertainty region

$$r_{\mathrm{MM}}(d(\cdot)) = \sup_{(\Lambda,\Gamma)\in\eta(\hat{\Lambda},\hat{\Gamma})} \sum_{W\in\Omega_W} P_\Gamma(W)\int_{X\in\Omega_X} l_{01}(W,d(X))P_\Lambda(X|W)dX . \tag{4}$$

Minimax decision rule has been derived as [8]

$$d_{\mathrm{MM}}(X) = \arg\max_W \left[P_{\hat{\Gamma}}(W) \max_{\Lambda\in\eta(\hat{\Lambda})} P_\Lambda(X|W) \right], \tag{5}$$

where HMM parameters were searched around neighborhood $\eta(\hat{\Lambda})$ by ML approach. Also, BPC is used to compensate model variations by averaging the uncertainties of Λ,Γ expressed by the prior densities $P(\Lambda),P(\Gamma)$. The overall risk of a BPC decision is given by

$$\begin{aligned} r_{\mathrm{BPC}}(d(\cdot)) &= E_{(W,X)}E_{(\Lambda,\Gamma)}[l_{01}(W,d(X))] \\ &= \sum_{W\in\Omega_W}\int_{X\in\Omega_X} \left[\begin{array}{l} l_{01}(W,d(X))\int_{\Omega_\Lambda} P_\Lambda(X|W)p(\Lambda)d\Lambda \\ \times \int_{\Omega_\Gamma} P_\Gamma(W)p(\Gamma)d\Gamma \end{array} \right] dX . \end{aligned} \tag{6}$$

where the integrals in bracket are known as *predictive distributions* $\tilde{P}(X|W)$ and $\tilde{P}(W)$ serving as *distribution estimates* for acoustic and language models. BPC decision was simplified to [6]

$$d_{\mathrm{BPC}}(X) = \arg\max_W P_{\hat{\Gamma}}(W)\int_{\Omega_\Lambda} P(X|W,\Lambda)p(\Lambda)d\Lambda . \tag{7}$$

In minimax and BPC decisions, only a zero-one loss function and the uncertainty of HMM were considered. These decisions dealt with the same assumption but used different distortion models.

2.2. GMBR Classification

Considering the third assumption, we introduce an adaptive loss function instead of zero-one loss function. Typically, the nonnegative real-valued loss function $l(W,d(X))$ should reflect the actual cost induced by a misclassification $d(X)$ of test sentence X with the target transcription W. To obtain a metric evaluating word error, we can measure the cost in word segments of test sentence in online unsupervised mode. The higher the word error rate, the larger the penalty assigned to measure the Bayes risk. Goel's minimum Bayes risk (GMBR) decision [5] uses the WER loss function $l_{\mathrm{WER}}(W,d(X))$ given by the Levenshtein distance between target W and hypothesis $d(X)$ strings

$$d_{\mathrm{GMBR}}(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} l_{\mathrm{WER}}(W,d(X))P_{\hat{\Lambda},\hat{\Gamma}}(W|X) . \tag{8}$$

Acoustic and language parameters $\hat{\Lambda},\hat{\Gamma}$ were assumed to be accurate for approximating the word posterior probability [5]

$$P_{\hat{\Lambda},\hat{\Gamma}}(W|X) = \frac{P_{\hat{\Lambda}}(X|W)P_{\hat{\Gamma}}(W)}{\sum_{W'} P_{\hat{\Lambda}}(X|W')P_{\hat{\Gamma}}(W')} . \tag{9}$$

N-best list or word lattice from the recognizer serves as the hypotheses $\{W'|W'\neq W\}$. Bayes risk of GMBR decision becomes

$$r_{\mathrm{GMBR}}(d(\cdot)) = E_{(W,X)}[l_{\mathrm{WER}}(W,d(X))] . \tag{10}$$

Typically, no probabilistic models and parameters are considered in WER loss function $l_{\mathrm{WER}}(W,d(X))$.

3. PMBR Classification Rule

Owing to these assumptions in Bayes classification, we should compensate the environmental mismatch and the ill-posed model in calculation of Bayes risk. For example, in noisy speech recognition, we are lacking for the prior information of noise type, signal-to-noise ratio and speaker features. We don't know the true distributions of $P_\Lambda(X|W),P_\Gamma(W)$ from insufficient data X and unreliable transcription W. We should simultaneously trace the *mismatch sources* in data W,X and characterize the *variability of the estimated parameters* $\hat{\Lambda},\hat{\Gamma}$. Model generalization can be assured to elevate classification performance of test sentence. For this concern, the Bayes risk should be not only averaged over the randomness of W,X but also Λ,Γ. We try to fulfill Bayes classification through minimizing the predictive Bayes risk

$$r_{\mathrm{PMBR}}(d(\cdot)) = E_{(W,X)}E_{(\Lambda,\Gamma)}[l_{\mathrm{BF}}(W,d(X))] . \tag{11}$$

The predictive minimum Bayes risk (PMBR) classification is developed through combining BPC and GMBR decisions where the uncertainty of Λ,Γ and the adaptive loss function are merged

$$d_{\mathrm{PMBR}}(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} l_{\mathrm{BF}}(W,d(X))\,\tilde{P}(W|X) . \tag{12}$$

We present a Bayes loss function $l_{\mathrm{BF}}(W,d(X))$ and a predictive word posterior probability $\tilde{P}(W|X)$ for Bayes classification.

3.1. Bayes Loss Function

Different from GMBR using the WER loss function, we present a statistical loss function through solving a hypothesis-test problem. The loss due to the classification action $d(X)$ is formulated as a *confidence measure* towards accepting the null hypothesis H_0 occurring loss event against the alternative hypothesis H_1 occurring lossless event [2]. Beyond the likelihood ratio test, we cope with the test by measuring the Bayes factor [7]

$$\begin{aligned} b(W,d(X)) &= \frac{\tilde{P}(X,d(X)|H_0 : d(X)\neq W)}{\tilde{P}(X,d(X)|H_1 : d(X)=W)} \approx \\ &\frac{\sum_{(d(X)=W')\neq W}\int_{\Omega_\Lambda} P_\Lambda(X|W')P(\Lambda)d\Lambda\int_{\Omega_\Gamma} P_\Gamma(W')P(\Gamma)d\Gamma}{\int_{\Omega_\Lambda} P_\Lambda(X|W)P(\Lambda)d\Lambda\int_{\Omega_\Gamma} P_\Gamma(W)P(\Gamma)d\Gamma} . \end{aligned} \tag{13}$$

Null hypothesis is accepted if Bayes factor exceeds a critical threshold. Importantly, Bayesian approach assumes randomness of

parameters Λ,Γ so that the predictive distributions $\tilde{P}(X|W)$ and $\tilde{P}(W)$ are seen in (13). Bayes factor offers a way of incorporating external information for penalizing misclassifications of speech signals. The numerator sums up joint predictive distributions $\tilde{P}(X,d(X))$ corresponding to wrong classification $d(X)\neq W$, whereas the denominator involves only the predictive distribution for correct classification $d(X)=W$. The numerator can be approximated using non-target strings $\{(d(X)=W')\neq W\}$ provided by classifier $d(X)$. Empirically, we smooth the logarithm of Bayes factor via a sigmoid function and establish the Bayes loss function by

$$l_{\mathrm{BF}}(W,d(X))=\frac{1}{1+\exp(-\gamma\log b(W,d(X))+\theta)}. \quad (14)$$

where γ and θ are variables tuning the degree of nonlinearity. This Bayes loss is computed as a *continuous* value, which is related to discrete-valued WER. The higher the Bayes loss, the more likely the misclassification increases the WER.

3.2. Predictive Word Posterior Probability

In this paper, the Bayesian treatment is not only taken in Bayes loss function but also in word posterior probability. We calculate the predictive word posterior probability by

$$\tilde{P}(W|X)=\frac{\tilde{P}(X|W)\tilde{P}(W)}{\tilde{P}(X)}\approx\frac{\tilde{P}(X|W)\tilde{P}(W)}{\sum_{W'}\tilde{P}(X|W')\tilde{P}(W')}$$
$$=\frac{\int_{\Omega_\Lambda}P_\Lambda(X|W)P(\Lambda)d\Lambda\int_{\Omega_\Gamma}P_\Gamma(W)P(\Gamma)d\Gamma}{\sum_{W'}\int_{\Omega_\Lambda}P_\Lambda(X|W')P(\Lambda)d\Lambda\int_{\Omega_\Gamma}P_\Gamma(W')P(\Gamma)d\Gamma}. \quad (15)$$

This predictive word posterior probability is determined by replacing the point estimates of distributions $P_\Lambda(X|W),P_\Gamma(W)$ in word posterior probability of (9) with the predictive distributions $\tilde{P}(X|W),\tilde{P}(W)$. The uncertainties of parameters Λ,Γ are normalized. In comparison of (13) and (15), it is interesting that Bayes factor equals the inverse of word posterior probability if the same hypothesis set $\{W'\}$ is used to approximate the predictive distribution $\tilde{P}(X,d(X)|H_0)$ and the evidence term $\tilde{P}(X)$ even though two distributions are inherently different. The Bayes loss function in PMBR classification can be interpreted as an additional smoothing of the predictive posterior probability $\tilde{P}(W|X)$. The higher the predictive word posterior probability is calculated, the lower the Bayes loss is measured as a smoothing factor to determine the Bayes risk. In implementation of PMBR, the Bayes loss is calculated for *N-best list rescoring*.

3.3. Predictive Distribution

In the evaluation, we only investigate the uncertainties of HMM mean vectors $\Lambda=\{\mu_{ik}\}$ for different states i and mixture components k. The remaining HMM parameters including initial state probabilities $\{\pi_i\}$, state transition probabilities $\{a_{ij}\}$, mixture weights $\{\omega_{ik}\}$ and covariance matrices $\{\Sigma_{ik}\}$ are assumed to be deterministic in calculation of Bayes risk. A good way to model the uncertainty of a Gaussian mean vector is to use the *conjugate prior*, which is a Gaussian density $P(\mu_{ik}|\varphi_{ik})=N(\mu_{ik};m_{ik},C_{ik})$ with mean vector m_{ik} and covariance matrix C_{ik}. The hyperparameters $\varphi=\{\varphi_{ik}\}=\{m_{ik},C_{ik}\}$ should sufficiently reflect the randomness of parameters $\{\mu_{ik}\}$. The predictive distribution of speech frame $\mathbf{x}_t$ with word w, given a random mean vector μ_{ik} and the deterministic covariance matrix $\hat{\Sigma}_{ik}$, is expressed by

$$\tilde{P}(\mathbf{x}_t|w)=\int P(\mathbf{x}_t|\mu_{ik},\hat{\Sigma}_{ik})P(\mu_{ik})d\mu_{ik}\sim N(\mathbf{x}_t;m_{ik},\hat{\Sigma}_{ik}+C_{ik}), \quad (16)$$

which is a closed-form integral as a Gaussian distribution. The predictive distribution of a whole sentence $X=\{\mathbf{x}_t\}$ can be determined by the Viterbi approximation

$$\tilde{P}(X|W)=\int_{\Omega_\Lambda}P_\Lambda(X|W)P(\Lambda)d\Lambda\approx\int_{\Omega_\Lambda}P_\Lambda(X,\hat{\mathbf{s}},\hat{\mathbf{l}}|W)P(\Lambda)d\Lambda$$
$$\approx\hat{\pi}_{\hat{s}_1}\prod_t\hat{a}_{\hat{s}_t\hat{s}_{t+1}}\hat{\omega}_{\hat{s}_t\hat{l}_t}N(\mathbf{x}_t;m_{\hat{s}_t\hat{l}_t},\hat{\Sigma}_{\hat{s}_t\hat{l}_t}+C_{\hat{s}_t\hat{l}_t}), \quad (17)$$

where the optimal state and mixture component sequences $\hat{\mathbf{s}}=\{\hat{s}_t\},\hat{\mathbf{l}}=\{\hat{l}_t\}$ are merged for computing acoustic score.

4. Experiments

4.1. Databases and Experimental Setup

To evaluate the robustness of Bayes classification rules, we conducted speech recognition of connected Chinese digits in car environments [1]. Two databases were prepared. The first database was recorded in office environments via close-talking microphones. There were 1000 utterances of connected digits from 50 males and 50 females. Each speaker uttered ten sentences. We applied these utterances to train speaker-independent (SI) HMMs. Also, we collected another test database CARNAV98 containing utterances of five males and five females different from those in training data and recorded in two medium class cars. These utterances were collected using a hands-free far-talking microphone. We had three sessions of *standby*, *downtown* and *freeway* conditions with the averaged car speeds being 0, 50 and 90 km/h and the averaged signal-to-noise ratios being 8.0, -3.1 and -7.0 dB, respectively. During recording, we kept the engine on, the air-conditioner on, the music off and the windows rolled up. The numbers of test utterances were 50, 150 and 250 for standby, downtown and freeway conditions, respectively. Each speaker provided five adaptation utterances (N=5) for estimating $\{m_{ik},C_{ik}\}$. WERs were averaged over ten test speakers. All utterances contained three to eleven random digits. We modeled each Chinese digit using a left-to-right seven-state HMM without state skipping. There were 73 HMM states (70 for ten digits, 1 for pre-silence, 1 for post-silence and 1 for silence within connected digits). Each HMM state was composed of four mixture components. A speech frame was characterized by a feature vector with 12 LPC-derived cepstral coefficients, 12 delta cepstral coefficients, one delta log energy and one delta delta log energy. No language model $P_\Gamma(W)$ was involved. In the experiments, we compared MAP, minimax, BPC, GMBR and PMBR decision rules. MAP decoding with SI HMMs was referred as the baseline system. Using GMBR, we performed MAP adaptation [4] and used the adapted parameters for decoding the test data.

4.2. Implementation Issues

In BPC and PMBR, the hyperparameters $\varphi=\{m_{ik},C_{ik}\}$ were empirically estimated from training and adaptation data. To do so, we individually estimated HMM mean vectors $\{\hat{\mu}_{ik}^s\}$ for each training speaker s. The hyperparameters m_{ik} and C_{ik} were determined by calculating the sample mean vector and the sample

covariance matrix using the estimated means $\{\hat{\mu}_{ik}^{s}\}$ over speakers $s = 1, \cdots, S$. The estimated hyperparameters modeled *cross speaker variability*. As shown in (16)(17), the hyperparameter C_{ik} was added to the HMM covariance matrix $\hat{\Sigma}_{ik}$, which modeled *total variability*. It was meaningful that the variance in predictive distribution could represent the inter-speaker variability. To capture intra-speaker variability, we further used adaptation data and performed MAP adaptation [4] of hyperparameter m_{ik} to target speaker. Hyperparameter C_{ik} was unchanged. Providing the adaptation data and the marginalization in (16)(17), we established a predictive decision rule for unknown test data. Similar to GMBR decision [5], we used an exponential discounting weight α to balance the scores between loss function and word posterior probability. PMBR decision was made according to the criterion

$$\sum_{W \in \Omega_W} l_{\mathrm{BF}}(W, d(X))^{\alpha} \tilde{P}(W|X). \qquad (18)$$

Word-level Bayes losses and predictive word posterior probabilities were calculated for individual word segments using the confusion sets obtained in lattice alignment procedure. N-best list was rescored accordingly. At each position in the alignment we picked up the best word hypothesis with the lowest Bayes risk. In the experiments, we selected $\alpha = 1.6$ for GMBR and $\alpha = 1.2$ for PMBR. Sigmoid parameters were set to $\gamma = 0.8$ and $\theta = 0.2$.

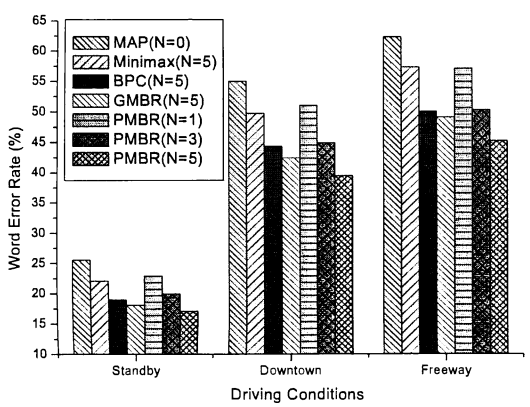

Figure 1: Comparison of WERs using different classification rules.

4.3. Experimental Results

In the experiments, MAP decision reports WERs of 25.6%, 55% and 62.3% for standby, downtown and freeway conditions, respectively [1]. Figure 1 displays WERs of classifiers using MAP, minimax, BPC, GMBR and PMBR with N=1, N=3 and N=5 in different driving conditions. The distortion model using BPC is better than that using minimax classifier. The hyperparameters m_{ik} and C_{ik} provide informative statistics of target speaker and noise. These two classifiers outperform MAP classifier with no distortion model involved. However, we find that GMBR and PMBR reduce WERs compared to MAP, BPC and minimax decisions due to the incorporation of loss function and word predictive probability. The performance of BPC gets close to that of GMBR. PMBR improves WERs by increasing adaptation data N and obtains the lowest WERs among these classifiers for three driving conditions. This reveals the superiority of using predictive distributions in noisy speech recognition. The higher the driving speed is involved, the more the reduction of word error rate is attained. In case of downtown condition, PMBR achieves WER 39.5%, which is significantly better than 49.7% using minimax, 44.3% using BPC and 42.4% using GMBR. The predictive distributions do help putting the environmental statistics into integration of Bayes risk and elevating MBR based speech recognition in car environments.

5. Summary

To pursue the essence of Bayesian theory for speech recognition, we introduced MAP, minimax, BPC and GMBR rules and proposed a new PMBR rule to compensate the weaknesses of decoding algorithms without considering the randomness of HMM parameters and the statistical representation of loss function. Through testing classification loss of speech signal, we presented Bayes factor to develop Bayes loss function for speech recognition. Also, we used the predictive distributions in calculation of word posterior probability, or equivalently the predictive Bayes risk. The prior information of HMM mean parameters was merged so that the decision rule was robust to variations of parameter estimation. In experiments on a car noisy speech database, we investigated the effects of hyperparameters in Bayes decision rules. We showed the superiority, in terms of WER, of speech recognition using PMBR. In the future we will continue exploring the effects of uncertainties of other HMM parameters and *n*-gram parameters.

6. References

[1] J.-T. Chien and G.-H. Liao, "Transformation-based Bayesian predictive classification using online prior evolution", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, pp. 399-410, 2001.

[2] J.-T. Chien, C.-H. Huang, K. Shinoda and S. Furui, "Towards optimal Bayes decision for speech recognition", in *Proc. ICASSP*, pp. 45-48, 2006.

[3] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities", in *Proc. ICASSP*, pp. 1655-1658, 2000.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 291–298, 1994.

[5] V. Goel and W. Byrne and S. Khudanpur, "LVCSR rescoring with modified loss function: a decision theoretic perspective", in *Proc. ICASSP*, pp. 425-428, 1998.

[6] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2, pp. 200-204, 2000.

[7] R. E. Kass and A. E. Raftery, "Bayes factors", *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773-795, 1995.

[8] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 1, pp. 90-100, 1993.

[9] L. Mangu, E. Brill and A. Stolcke, "Finding consensus among words: lattice-based word error minimization", in *Proc. EUROSPEECH*, pp. 495-498, 1999.

[10] F. Wessel, R. Schluter and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities", in *Proc. ICASSP*, pp. 33-36, 2001.