論文 / 著書情報
Article / Book Information

Title	Towards Optimal Bayes Decision for Speech Recognition
Author	Jen-Tzung Chien, Chin-Hsien Huang, Koichi Shinoda, Sadaoki Furui
Journal/Book name	Proc. ICASSP2006, Vol. , No. , pp. SLP-L2.6
発行日 / Issue date	2006, 5
権利情報 / Copyright	(c)2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Towards Optimal Bayes Decision for Speech Recognition

Jen-Tzung Chien[a], *Chih-Hsien Huang*[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
{chien, chhuang}@chien.csie.ncku.edu.tw & {shinoda, furui}@cs.titech.ac.jp

ABSTRACT

This paper presents a new speech recognition framework towards fulfilling optimal Bayes decision theory, which is essential for general pattern recognition. The recognition procedure is developed through minimizing the Bayes risk, or equivalently the expected loss due to classification action. Typically, loss function measures the penalty/evidence of choosing a candidate hypothesis. This function was manually specified or empirically calculated. Here, we exploit a novel *Bayes loss function* via testing the hypotheses whether the classification action produces loss or not. A Bayes factor is derived to measure loss in a statistical and meaningful way. Attractively, Bayes loss function using predictive distributions is robust to the uncertainty of environments. Also, optimizing this Bayes criterion equals to minimizing classification errors of test data. The relation between the minimum classification error (MCE) classifier and the proposed optimal Bayes classifier (OBC) is bridged. Specifically, the logarithm of Bayes factor in OBC is analogous to the misclassification measure in MCE when using predictive distribution as the discriminant function. We accordingly build a robust and discriminative classification for large vocabulary continuous speech recognition. In the experiments on broadcast news transcription, the new OBC rule significantly outperforms traditional maximum *a posteriori* classification.

1. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) plays a crucial role of establishing many human-machine communication systems. Great research efforts have been made to solve different issues in LVCSR system. Among these issues, how to build a robust classification rule or decoding algorithm is critical to achieve desirable speech recognition performance. Starting from statistical pattern recognition (PR) [3] viewpoint, we should tackle this issue following the optimal Bayes decision theory where the Bayes risk or expected loss is minimized to achieve desirable LVCSR performance. Accordingly, Goel and Byrne [5] developed the minimum Bayes risk (MBR) classification to improve conventional speech recognition paradigm selecting the word string with the highest posterior probability or equivalently adopting the maximum *a posteriori* (MAP) decision rule. In fact, MAP classification was a special realization of MBR classification when assigning equal loss for different misclassification actions. Nevertheless, loss function acts as a classification penalty, which is measurable for different misclassifications from test data. If we can incorporate representative loss functions, MBR should be much better than MAP classification. Conceptually, loss function has similar meaning of merging confidence measure in decoding algorithm [16]. Word posterior probability was merged to improve LVCSR. Also, confidence based search algorithm was exploited to prune unlikely hypotheses using likelihood ratio test [1].

Intuitively, MBR criterion is feasible to build optimal Bayes classification procedure although it was also applied for hidden Markov model (HMM) training [10]. When surveying loss function in MBR classification, it was popular to use word error rate loss function measured by Levenshtein distance matching two strings in a fashion of dynamic programming [6]. Attractively, implementing MBR criterion was equivalent to finding a classifier with minimum word error rate. Word error minimization using N-best list [14] and lattice [13] rescoring was implemented. Also, Wessel *et al.* [15] presented the time frame error rate loss function for fulfilling MBR classification. In general, the previous loss functions were deterministic and representative for measuring similarity between target string and hypothesis string. To realize a truly optimal Bayes decision for speech recognition, we present a novel *Bayes loss function* determined in statistical Bayesian viewpoint. We test the hypothesis of producing loss caused by a classification action. A Bayes factor [8][11] is accordingly calculated for loss measurement using the log predictive distribution ratio of null and alternative hypotheses. To distinguish different classifiers, here, MBR classification using Bayes loss function is referred as the *optimal Bayes classification* (OBC) because this classifier is carried out towards obeying optimal Bayes theories. Interestingly, we connect the relation between minimum classification error (MCE) [9] and OBC classifiers. In this manner, we illustrate the discriminative capability using OBC decision rule. From the experiments using Mandarin broadcast news speech database, we find that the proposed OBC obtains better LVCSR performance compared to MAP and conventional MBR classification.

2. SURVEY OF DECISION RULES

According to statistical PR theory, optimal Bayes decision rule is established through minimizing Bayes risk or expected loss. For the application of speech recognition, we minimize the expectation of loss associated with a decision $d(X)$ if the true transcription of speech signal $X \in \Omega_X$ is $W \in \Omega_W$. Bayes risk is expressed by

$$E_{(W,X)}[l(W,d(X))] = \int_{X\in\Omega_X} P(X)\left[\sum_{W\in\Omega_W} l(W,d(X))P(W|X)\right]dX \quad (1)$$
$$= \sum_{W\in\Omega_W} P_\Gamma(W)\int_{X\in\Omega_X} l(W,d(X))P_\Lambda(X|W)dX \equiv r(d(\cdot))$$

OBC decision rule is constructed by minimizing overall risk, namely

$$d_{\mathrm{OBC}}(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} l(W,d(X))P(W|X)$$
$$= \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} P_\Gamma(W)\int_{X\in\Omega_X} l(W,d(X))P_\Lambda(X|W)dX \quad (2)$$

In real-world implementation, we assume that (1) observation space Ω_X is known, (2) true distributions of acoustic model $P_\Lambda(X|W)$ and language model $P_\Gamma(W)$ are known and (3) loss function $l(W,d(X))$ is given [12]. The resulting speech recognition performance is limited accordingly. To deal with the first issue, we can build the adaptive decision rule via decision parameter adaptation. Through adaptation of HMM Λ and n-gram Γ parameters, decision rule is adaptive to the unknown observation space. Also, when considering the second issue, the Bayesian predictive classification (BPC) [2][7] was exploited to achieve a robust decision where the uncertainties of parameters described by prior distributions $P(\Lambda)$ and $P(\Gamma)$ were averaged as follows

$$d_{\mathrm{BPC}}(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} \int_{X\in\Omega_X} \left[\begin{array}{l} l(W,d(X))\int_{\Omega_\Lambda} P_\Lambda(X|W)p(\Lambda)d\Lambda \\ \times \int_{\Omega_\Gamma} P_\Gamma(W)p(\Gamma)d\Gamma \end{array} \right] dX \quad (3)$$

This BPC decision rule was better than conventional method where the estimated Λ and Γ were pretended to be true distribution parameters. In (3), the integrals in bracket $\tilde{P}(X|W)$ and $\tilde{P}(W)$ are predictive distributions corresponding to acoustic and linguistic models. In this paper, we concern the third issue and present a novel loss function for OBC speech recognition. This serves as the most important issue to accomplish MBR or OBC decision rule.

Traditionally, speech recognition systems adopt the empirical loss function determined from training data or simply use the zero-one loss function

$$l(W,d(X)) = \begin{cases} 0, & \text{if } W = d(X) \\ 1, & \text{if } W \neq d(X) \end{cases}. \quad (4)$$

Namely, correct classification has no loss while different wrong classification is penalized with equal loss. For this case, OBC decision rule is reduced to MAP decision rule

$$d_{\mathrm{MAP}}(X) = \hat{W} = \arg\max_W P(W|X) = \arg\max_W P_\Lambda(X|W)P_\Gamma(W), \quad (5)$$

where the posterior distribution $P(W|X)$ is maximized to find optimal word sequence $\hat{W}$. Equivalently, we search the optimal solution with the highest acoustic $P_\Lambda(X|W)$ and linguistic $P_\Gamma(W)$ scores. However, the real-valued loss function $l(W,d(X))$ should properly reflect the cost induced when a test utterance X with true transcription W is recognized as $d(X)$. This cost can be measured from word segments of test data in an online unsupervised mode. Generally, the higher the word error rate is measured for a classification pair $(W,d(X))$, the larger the penalty/cost should be assigned for accumulating Bayes risk. In [5][6], the word error rate (WER) loss function $l_{\mathrm{WER}}(W,d(X))$ calculated by Levenshtein distance was developed for building MBR classification

$$d_{\mathrm{MBR}}(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} P_\Gamma(W)\int_{X\in\Omega_X} l_{\mathrm{WER}}(W,d(X))P_\Lambda(X|W)dX. \quad (6)$$

Using WER loss function, N-best list or word lattice was rescored so as to improve LVCSR performance. Basically, previous loss functions were determined via performing dynamic programming of two hypothesis strings. No probabilistic models and parameters were considered in loss function calculation. To build a truly Bayesian framework, we are motivated to present a Bayes loss function for OBC decision. We are measuring the probabilistic similarity between target and hypothesis strings and coming up with a new OBC decision rule completely constructed using Bayes theory.

3. OPTIMAL BAYES CLASSIFICATION

Our goal aims at fulfilling optimal Bayes decision for LVCSR. Different from MBR speech recognition [5][6], we present a statistical loss function derived from hypothesis test theory using Bayesian approach. We are describing the robust and discriminative capabilities of applying proposed OBC decision rule.

3.1. Test of Classification Loss

The specification of loss function is crucial for optimal Bayes decision. Basically, whether the classification $d(X)$ produces loss or not is referred as a two-class PR problem. To formulate the *confidence* or loss due to a classification action, we describe the mathematical model as a hypothesis test problem. According to the outcomes of loss and lossless classification events, null H_0 and alternative H_1 hypotheses are naturally defined by

H_0: test data X is misclassified, or $d(X)$ produces loss.

H_1: test data X is not misclassified, or $d(X)$ is lossless.

From Neyman-Pearson's Lemma, the optimal solution to hypothesis testing is called likelihood ratio test. Having probability distributions of two hypotheses, null hypothesis H_0 is accepted if likelihood ratio exceeds a critical threshold

$$\mathrm{LR} = \frac{P(X,d(X)|H_0 : d(X)\neq W)}{P(X,d(X)|H_1 : d(X)=W)} > \tau. \quad (7)$$

However, we don't know true distributions of null hypothesis $P(X,d(X)|H_0)$ and alternative hypothesis $P(X,d(X)|H_1)$. Implementation of loss function using likelihood ratio $l_{\mathrm{LR}}(W,d(X))$ shall be sensitive to the uncertainties of distribution forms and trained parameters. To setup a loss function robust to uncertainties of speech models, we present a novel Bayes loss function for OBC speech recognition. This function is built by solving hypothesis test problem using Bayesian approach [8][11] where model parameters $\Xi = \{\Lambda,\Gamma\}$ are random with prior distributions $P(\Lambda)$ and $P(\Gamma)$. A Bayes factor is yielded and expressed using predictive distributions $\tilde{P}(X|W)$ and $\tilde{P}(W)$

$$\mathrm{BF} = b(W,d(X)) = \frac{\tilde{P}(X,d(X)|H_0)}{\tilde{P}(X,d(X)|H_1)} = \frac{\sum_{d(X)\neq W}\tilde{P}(X|d(X))\tilde{P}(d(X))}{\tilde{P}(X|W)\tilde{P}(W)}$$
$$= \frac{\sum_{d(X)\neq W}\int_{\Omega_\Lambda} P_\Lambda(X|d(X))P(\Lambda)d\Lambda\int_{\Omega_\Gamma} P_\Gamma(d(X))P(\Gamma)d\Gamma}{\int_{\Omega_\Lambda} P_\Lambda(X|W)p(\Lambda)d\Lambda\int_{\Omega_\Gamma} P_\Gamma(W)P(\Gamma)d\Gamma} \quad (8)$$

In (8), the numerator sums up all joint predictive distributions $\tilde{P}(X,d(X))$ corresponding to misclassification actions $d(X)\neq W$ while the denominator involves only the predictive

distribution for the case of correct classification $d(X) = W$. In LVCSR implementation, the word candidate with the highest predictive score is referred as true transcription W. The other competing word candidates at the same word segment are included in word set of null hypothesis $d(X) \neq W$. These word candidates are found from word lattices produced by word graph generation algorithm. To determine an effective Bayes factor, we can empirically merge tuning factors when calculating predictive distributions for different competing words. Using this Bayes factor, we are able to develop a new parametric loss function better than conventional zero-one, confidence measure and word error rate loss functions. The resulting loss function is robust because predictive distributions are calculated to tackle randomness of trained model parameters.

3.2. Bayes Loss Function

In this study, word-level Bayes factor $b(W, d(X))$ is further normalized to build Bayes loss function. Our considerations are twofold. First, loss function is involved in combining acoustic $P_\Lambda(X|W)$ and linguistic $P_\Gamma(W)$ scores and rescoring the word lattices to obtain optimal word sequence for LVCSR. The value of loss function should be in a limited range to prevent deteriorating system performance. For the second reason, we are generating a perceptually meaningful loss function to connect the relation to MCE classifier. Therefore, we calculate the logarithmic Bayes factor and smooth the value in a range between zero and one using the sigmoid function, which is the most common form of activation function in construction of neural network. The Bayes loss function is formed by

$$l_{\mathrm{BF}}(W, d(X)) = l(b(W, d(X))) = \frac{1}{1 + \exp(-\gamma \log b(W, d(X)) + \theta)} \quad (9)$$

where γ and θ are two tuning parameters balancing the linearity and nonlinearity of activation function. Using this Bayes loss function, new OBC decision rule minimizing the expected Bayes loss function is established by

$$d_{\mathrm{OBC}}(X) = \arg\min_{d(X) \in \Omega_W} \sum_{W \in \Omega_W} l_{\mathrm{BF}}(W, d(X))\, P(W|X) . \quad (10)$$

We are finding the candidate word sequence $d(X) = W'$ producing the smallest Bayes risk. In what follows, we are illustrating the discriminability of using new OBC through connecting the relation to MCE classifier.

3.3. Relation between OBC and MCE Classifiers

MCE is a popular model training approach for estimating discriminative acoustic models Λ [9]. This approach assumed that the model parameters estimated from training data Y were fitted for recognizing unknown test data X. Given a specified discriminant function $g(Y, \Xi)$, the misclassification measure in MCE criterion is defined by

$$m(W, d(Y)) = -g(Y, \Xi_W) + \log\left[\sum_{d(Y) \neq W} \exp[g(Y, \Xi_{d(Y)})\eta]\right]^{1/\eta} \quad (11)$$

This measure is substituted into sigmoid function of (9) to obtain loss function $l_{\mathrm{MCE}}(W, d(Y); \Xi) = l(m(W, d(Y)))$. MCE aims to fulfill optimal Bayes decision and estimate speech parameters via minimizing the expected loss

$$\hat{\Xi} = \arg\min_{\Xi} E_Y[l_{\mathrm{MCE}}(W, d(Y); \Xi)] . \quad (12)$$

Notably, MCE minimizes the expected loss $E_Y[l_{\mathrm{MCE}}(W, d(Y); \Xi)]$ of training data Y for model training while OBC minimizes the expected Bayes loss $E_{(W,X)}[l_{\mathrm{BF}}(W, d(X))]$ of test data X for building decision rule. In MCE, the expectation is done with respect to acoustic variable. But, in OBC, the expectation is performed with respect to acoustic and linguistic variables. Interestingly, if we apply MCE criterion in *test phase* and set discriminant function as a *logarithm of predictive distribution*

$$\begin{aligned} g(X, \Xi_W) &= \log \tilde{P}(X, d(X) = W) = \log \int_{\Omega_\Xi} P_\Xi(X, W) P(\Xi) d\Xi \\ &= \log \int_{\Omega_\Lambda} P_\Lambda(X|W) p(\Lambda) d\Lambda + \log \int_{\Omega_\Gamma} P_\Gamma(W) P(\Gamma) d\Gamma \end{aligned} , \quad (13)$$

we investigate that the *logarithmic Bayes factor* in OBC is equivalent to the *misclassification measure* with $\eta = 1$ in MCE as shown by

$$\begin{aligned} \log b(W, d(X)) &= -\log \tilde{P}(X, W) + \log \sum_{d(X) \neq W} \tilde{P}(X, d(X)) \\ &= -\log \int_{\Omega_\Xi} P_\Xi(X, W) P(\Xi) d\Xi + \log \sum_{d(X) \neq W} \int_{\Omega_\Xi} P(X, d(X)) P(\Xi) d\Xi . \\ &= m(W, d(X)) \end{aligned} \quad (14)$$

Attractively, we can interpret misclassification measure in MCE as a logarithmic Bayes factor or *confidence measure* for testing the hypothesis of misclassification action against that of correct classification action. Therefore, considering these properties, it is meaningful to claim that OBC decision rule can achieve discriminative classification because minimizing expected loss for OBC is comparable to minimizing classification errors or enhancing model discriminability.

4. EXPERIMENTS

4.1. Databases and Experimental Setup

We carried out Bayes decision rules for broadcast news transcription. LVCSR decoder contained lexicon tree, acoustic model and language model. In lexicon set, we used 74,868 Chinese words. Each word had at most four characters. All words were organized in a tree structure for within-word search. Acoustic model set consisted of Initial/Final sub-syllable HMM's for Mandarin speech recognition. Initial and Final HMM's had three and five states, respectively. Each state had at most 32 Gaussian mixture components. We used 7080 utterances from TCC300 speech database to train seed speaker independent HMM's. Then, we performed MAP task adaptation for LVCSR of MATBN broadcast news corpus using 680 MATBN utterances (35.8 minutes). In order to estimate the parameters in sigmoid function, loss function and language model weighting, we prepared 700 utterances as the held-out set. Also, there were 500 test utterances (11 minutes) containing 4105 characters. MATBN were shared by the Public Television Service Foundation of Taiwan and collected by Academia Sinica, Taiwan. Each speech frame was parameterized as 39-dimensional feature vector of 12 Mel-frequency cepstral coefficients (MFCC), one log energy and their first and second derivatives. Sentence-based cepstral mean subtraction was performed. Trigram language model was trained using CIRB corpus (about 342MB) via SRI language model toolkit. Good-Turing smoothing was applied. Language model perplexity was 437. We reported character error rate (CER) performance for different decision rules. In this study, we compared MAP, MBR and OBC decision rules. Different loss functions were incorporated in word graph rescoring. Word-conditioned tree copy search was

performed to build word graph. MAP decoding was referred as the baseline system. MBR with Levenshtein loss function was implemented for comparison. OBC decoding using Bayes loss function calculated with predictive distribution and with different priors were investigated.

4.2. Implementation Issues

In OBC rule implementation, we only calculated predictive distribution $\tilde{P}(X|\mu,W)$ considering the uncertainty of HMM mean vector μ. The other HMM parameters and trigram parameters were assumed to be deterministic. Prior density of HMM mean vector was modeled by a state-level tied Gaussian distribution $P(\Lambda)=P(\mu)=N(\mu|m,\Sigma)$ with mean vector m and covariance matrix Σ. Predictive distribution was derived in a form of Gaussian distribution [3]. Here, hyperparameters (m,Σ) were empirically estimated from training data via taking sample mean and variance of maximum likelihood parameters. Similar technique was applied to determine those hyperparameters corresponding to competing words or null hypothesis $d(X)\neq W$. Also, we merged the exponential weighting factors in criterion

$$\sum_{W\in\Omega_W} l_{\mathrm{BF}}(W,d(X))^{\alpha}[(P_{\Lambda}(X|W)^{1/\beta}P_{\Gamma}(W))/\textstyle\sum_{W^{(c)}}P_{\Lambda}(X|W^{(c)})^{1/\beta}P_{\Gamma}(W^{(c)})] \quad (15)$$

We searched optimal factors in (9) and (15) using held-out data. Here, we fixed $\theta=0$ in (9) and $\beta=8$ in (15). By searching factors with minimum Bayes risk, we selected $\gamma=1$ and $\alpha=0.9$ in realization of OBC decoding and $\alpha=1.6$ for MBR decoding. In calculation of loss function using Bayes factor and Levenshtein distance, we considered at most twenty competing words for each word hypothesis. The lattice alignment procedure [13] was performed to determine word segments. Loss function was calculated within each word segment.

4.3. Experimental Results

Experimental results show that the baseline system using MAP decoding obtains CER of 39.8%. Using OBC, we investigated two kinds of priors to evaluate the effect on LVCSR performance. In PRIOR I, we performed MLLR adaptation of prior parameters from TCC300 corpus to broadcast news environments using task adaptation data. In PRIOR II, we further adapted these priors to the test speakers using held-out data. Using OBC with PRIOR I, we reduce CER to 38.2%. However, when adopting PRIOR II in OBC decision, CER is further reduced to 37.7% which is better than 38.4% using Levenshtein loss function based MBR. These results indicate that OBC decision rule outperforms MAP decision and MBR decision. This reveals the superiority of using OBC decoding on broadcast news transcription task. Importantly, if we use the priors closer to test speakers/environments, the performance could be improved accordingly. We will continue investigating the effects of hyperparameters and the predictive densities considering model uncertainties of other HMM and *n*-gram parameters.

5. SUMMARY

We have surveyed a series of decision rules and proposed new OBC decision rule for broadcast news transcription. Through hypothesis test of classification loss, we developed a Bayes factor to measure the Bayes loss for building optimal Bayes decision rule. Using sigmoid function, we derived Bayes loss function for OBC decision. It was a complete Bayesian solution to building decision rule for LVCSR. Attractively, this OBC rule was robust and discriminative due to the incorporation of prediction density and the relation to MCE classification. Using logarithm of predictive distribution as discriminant function, the logarithmic Bayes factor in OBC was equivalent to misclassification measure in MCE. Our experiments showed the superiority of OBC to other decision rules.

6. REFERENCES

[1] M. Afify, F. Liu, H. Jiang and O. Siohan, "A new verification-based fast match for large vocabulary continuous speech recognition", *IEEE Trans. Speech and Audio Processing*, vol.13, no. 4, pp. 546-553, 2005.

[2] 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.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

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

[5] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition", *Computer Speech and Language*, vol. 14, pp. 115-135, 2000.

[6] 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.

[7] 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.

[8] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: applications to speaker verification", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 8, pp. 874-884, 2001.

[9] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

[10] J. Kaiser, B. Horvat and Z. Kacic, "Overall risk criterion estimation of hidden Markov model parameters", *Speech Communication*, vol. 38, pp. 383-398, 2002.

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

[12] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241-1269, 2000.

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

[14] A. Stolcke, Y. Konig and M. Weintraub, "Explicit word error minimization in N-best list rescoring", in *Proc. EUROSPEECH*, pp. 163-165, 1997.

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

[16] F. Wessel, R. Schluter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 288-298, 2001.