ROBUST SPEECH RECOGNITION BASED ON VITERBI BAYESIAN PREDICTIVE CLASSIFICATION

Hui Jiang[†] *Keikichi Hirose*[†] *Qiang Huo*[‡]

[†]Department of Information and Communication Engineering, University of Tokyo, Japan
[‡]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

In this paper, we investigate a new *Bayesian predictive classification* (BPC) approach to realize robust speech recognition when there exist mismatches between training and test conditions but no accurate knowledge of the mismatch mechanism is available. A specific approximate BPC algorithm called Viterbi BPC (VBPC) is proposed for both isolated word and continuous speech recognition. The proposed VBPC algorithm is compared with conventional Viterbi decoding algorithm on speaker-independent isolated digit and connected digit string (TIDIGITS) recognition tasks. The experimental results show that VBPC can considerably improve robustness when mismatches exist between training and testing conditions.

1. INTRODUCTION

Although many advances have recently been achieved in automatic speech recognition, it is also found that the performance of a speech recognizer always degrades drastically whenever some acoustic mismatches between testing and training conditions exist. Today the dominant method to deal with these mismatches is to compensate the feature vectors or speech models in order to remove or reduce the mismatches between testing data and trained models (e.g. [1], [7], etc.). In the compensation approach, some prior knowledge about the mechanism of mismatches is necessary to design a suitable form of mapping function. In practice, we generally have no idea about the sources of variability in speech signal, and no full knowledge to figure out the mechanism of mismatches between training data in the laboratory and testing data in real field. In the extreme case, the only available information is the test data along with a set of pre-trained speech models. Some recent approaches have focused on modifying the decision rule and the model parameters so that part of the mismatch can be compensated and the decision performance can be improved. This scheme becomes a potential approach for robust speech recognition because it need not make rigid assumptions about sources of distortion. One such approach is the *minimax classification* algorithm [4] which assumes the best decision parameters for the given test data lie in the neighborhoods of the given parameters and adjusts the decision rule and the corresponding parameters accordingly. The minimax classification is thus geared to protect against the possibility of the worst mismatch. The main disadvantage of the minimax approach is that it usually do not perform nearly as well as in a less malign situation and/or those techniques which use some prior information of the possible mismatches. Another disadvantage is that it can not be extended easily to perform continuous speech recognition (CSR) because the combination of uncertainty neighborhoods surrounding the model parameters that need to be examined can become quite large[4, 6].

In this paper and [3], we investigate a new *Bayesian predictive classification* (BPC) approach to realize robust speech recognition when unknown mismatches exist between training and testing conditions. A specific approximate BPC algorithm called Viterbi BPC (VBPC) which aims at mitigating to some extent the above-mentioned difficulties of the minimax approach, is proposed in this paper. We apply the proposed VBPC framework to robust speaker independent recognition of Japanese isolated digits and TIDIGITS English connected digit strings under the mismatch caused by additive Gaussian white noise where each digit is modeled with a Gaussian mixture continuous density hidden Markov model (CDHMM). Our experimental results show that the proposed VBPC rule achieves considerable improvement in various signal-to-noise ratios (SNR) over the conventional Viterbi decoding in both tasks.

The remainder of the paper is organized as follows. In section 2, the basic principle of the BPC rule is briefly introduced. In section 3, the formulation of VBPC for CDHMM is presented. In section 4, a series of comparative experiments on isolated/connected digits recognition tasks are reported. Finally, our findings are summarized in section 5.

2. BAYESIAN PREDICTIVE CLASSIFICATION APPROACH

In a companion paper [3], we discuss how to apply the general BPC to CDHMM-based robust speech recognition, its relations to the conventional *plug-in maximum a posteriori* (MAP) decision rule and minimax decision rule, and finally focus on another approximate BPC method called *quasi-Bayesian predictive classification* (QBPC). In this paper, we focus our study on VBPC approach. Whenever possible, we use the same notations as those in [3].

Let's view a *word* W[1] and the associated acoustic observation $\mathbf{X}$ (usually, a feature vector sequence) as a jointly

[1] Depend on the problem of interest, *word* here could be any linguistic unit, such as a phoneme, a syllable, a word, a phrase, etc.

2

distributed random pair $(W, \mathbf{X})$. The proposed BPC rule [3] assumes some prior knowledge (albeit crude) about the *possible* mismatch, and at the same time takes into account its uncertainty in decision parameters. Only acoustic models are adjusted in this study. We use a prior PDF (probability density function) $p(\Lambda|\varphi)$ to represent our knowledge about the uncertainty of the unknown CDHMM parameters Λ (e.g. [2]). The BPC rule to obtain the recognition result, $\hat{W}$, is then

$$\hat{W} = \arg\max_{W} \tilde{p}(W|\mathbf{X}) = \arg\max_{W} \tilde{p}(\mathbf{X}|W) \cdot p(W) \qquad (1)$$

where

$$\tilde{p}(\mathbf{X}|W) = \int p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \qquad (2)$$

is called the predictive PDF of the observation $\mathbf{X}$ given the word W. The computation of this predictive PDF is the most difficult part of the BPC procedure.

3. VITERBI BAYESIAN PREDICTIVE CLASSIFICATION (VBPC)

In the CDHMM case, due to the nature of the *missing data* problem in HMM formulation (see related discussions in [2, 3]), it is not easy to compute the true *predictive density*:

$$\tilde{p}(\mathbf{X}|W) = \sum_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \qquad (3)$$

where $\mathbf{s}$ is the unobserved state sequence and $\mathbf{l}$ is the associated sequence of the unobserved mixture. component labels corresponding to the observation sequence $\mathbf{X}$. Consequently, some approximations are needed [3].

One way to compute the approximate predictive PDF is to use the following Viterbi approximation:

$$\tilde{p}(\mathbf{X}|W) \approx \max_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \qquad (4)$$

The resultant BPC rule is named as VBPC rule.

We adopt the same notations as those in [2] and denote an N-state CDHMM with parameter vector $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the state transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}$ for each state i with the state observation PDF being a mixture of multivariate Gaussians:

$$p(\mathbf{x}|\theta_i) = \sum\nolimits_{k=1}^{K} \omega_{ik} \mathcal{N}(\mathbf{x}|m_{ik}, r_{ik}), \qquad (5)$$

where the mixture coefficients ω_{ik}'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$, and $\mathcal{N}(\mathbf{x}|m_{ik}, r_{ik})$ is the k-th normal mixand with m_{ik} being the D-dimensional mean vector and r_{ik} being the $D \times D$ precision (inverse covariance) matrix. Given a test utterance $\mathbf{X} = (x_1, x_2, \cdots, x_T)$, the Viterbi Bayesian procedure for *approximately* [2] computing the approximate predictive PDF in equation (4) is described as follows:

(1) Initialization

$$\delta_1(i) = \tilde{\pi}_i \cdot \tilde{b}_i(x_1) \quad 1 \leq i \leq N \qquad (6)$$

$$\delta_1^0(i) = \tilde{\pi}_i \text{ and } \psi_1(i) = 0 \quad 1 \leq i \leq N \qquad (7)$$

where $\tilde{\pi}_i$ denotes the mean of the prior PDF of the HMM parameter π_i.

(2) Recursion: for $2 \leq t \leq T$, $1 \leq j \leq N$, do

$$\delta_t'(j) = \max_{1 \leq i \leq N} [\, \delta_{t-1}(i) \cdot \tilde{a}_{ij}'] \qquad (8)$$

$$\psi_t(j) = \arg\max_{1 \leq i \leq N} [\, \delta_{t-1}(i) \cdot \tilde{a}_{ij}'] \qquad (9)$$

where $\tilde{a}_{ij}'$ is the mean of the posterior PDF of the a_{ij} based on the optimal partial path up to the time instant t. i.e.

$$\tilde{a}_{ij}' = \begin{cases} \tilde{a}_{ij} & \text{for } i \neq j \\ \tilde{a}_{ij}^{L_j} / \tilde{a}_{ij}^{L_j - 1} & \text{for } i = j \end{cases} \qquad (10)$$

where $\tilde{a}_{ij}$ denotes the mean of the prior PDF of the HMM parameter a_{ij}, and $\tilde{a}_{ij}^{n}$ correspondingly denotes the nth order moment of a_{ij}.

if $j \neq \psi_t(j)$, then

$$\delta_t^0(j) = \delta_t'(j) \qquad (11)$$

$$\delta_t(j) = \delta_t'(j) \times \tilde{b}_j(x_t) \qquad (12)$$

else (i.e. $j = \psi_t(j)$)

$$\begin{aligned} \delta_t(j) \quad = \quad & \tilde{b}_j(x_{t-L_j+1}^{(s_j)(1)}, x_{t-L_j+2}^{(s_j)(2)}, \cdots, x_t^{(s_j)(L_j)}) \times \\ & \delta_{t-L_j+1}^0(j) \times \tilde{a}_{jj}^{L_j - 1} \end{aligned} \qquad (13)$$

where $x_t^{(s_i)(j)}$ means that x_t is the jth vector in the state i and L_j is the accumulated number of feature vectors belonging to state j based on the optimal partial path up to the time instant t.

(3) Termination

$$\tilde{p}(\mathbf{X}|W) \approx \max_i \; \delta_T(i) \qquad (14)$$

$$s_T^* = \arg\max_i \delta_T(i) \qquad (15)$$

(4) Path (state sequence) Backtracking

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \qquad t = T-1, T-2, \cdots, 1 \qquad (16)$$

The meaning of $\tilde{b}_j(\cdot)$ will be explained later.

In this paper, as the first step, we only consider the uncertainty of the mean vectors of CDHMM with diagonal covariance matrices and assume they are uniformly distributed in the neighborhoods of pre-trained means. The same uncertainty neighborhood shape as in [4] is adopted here:

$$\begin{aligned} \eta(\lambda) = \{ \lambda \mid & \pi_i = \pi_i^*, a_{ij} = a_{ij}^*, \omega_{ik} = \omega_{ik}^*, \\ & r_{ik} = r_{ik}^*, |m_{ikd} - m_{ikd}^*| \leq C\, d^{-1} \rho^d, \\ & 1 \leq i \leq N, 1 \leq k \leq K, 1 \leq d \leq D \} \end{aligned} \qquad (17)$$

[2] Strictly speaking, the following search algorithm can not completely warrant the eq.(4) in theory because the partial predictive value (i.e. δ_t) will possibly be re-computed partially in eq.(13) during search.

where hyperparameters C ($C > 0$) and ρ ($0 \leq \rho \leq 1$) are used to control respectively the possible mismatch *size* and *shape*, and $\{\pi_i^*, a_{ij}^*, m_{ihd}^*, r_{ih}^*\}$ denote the pre-trained model parameters.

We then have

$$\tilde{b}_i(x_t) \approx \omega_{il_t^*} \cdot \tilde{f}_{il_t^*}(x_t) = \omega_{il_t^*} \cdot \prod_{d=1}^{D} \tilde{f}_{il_t^* d}(x_{td}) \tag{18}$$

where l_t^* is the mixture component label to which x_t is "closest", and

$$\begin{aligned}
&\tilde{b}_i(x_{t-L_i+1}^{(s_i)(1)}, x_{t-L_i+2}^{(s_i)(2)}, \cdots, x_t^{(s_i)(L_i)}) \\
&= \prod_{k=1}^{K} \omega_{ik}^{L'_k} \cdot \tilde{f}_{ik}(x_{l_1^k}, \cdots x_{l_{L'_k}^k}) \\
&= \prod_{k=1}^{K} \omega_{ik}^{L'_k} \cdot \prod_{d=1}^{D} \tilde{f}_{ikd}(x_{l_1^k d}, \cdots x_{l_{L'_k}^k d})
\end{aligned} \tag{19}$$

where $x_{t-L_i+1}^{(s_i)(1)}, x_{t-L_i+2}^{(s_i)(2)}, \cdots, x_t^{(s_i)(L_i)}$ denote feature vectors belonging to state i in **X**, among which $l_1^k \cdots l_{L'_k}^k$ denote labels of the vectors "closest" to the mixture component k of state i. Then, with m_{ikd}^* and r_{ikd}^* being the pre-trained mean and precision parameters respectively, we have

$$\begin{aligned}
\tilde{f}_{ikd}(x_{1d}, x_{2d}, \cdots, x_{\zeta d}) &= (\frac{r_{ikd}^*}{2\pi})^{\frac{\zeta}{2}} \cdot \frac{1}{2Cd^{-1}\rho^d} \cdot \\
&\int_{m_{ikd}^* - Cd^{-1}\rho^d}^{m_{ikd}^* + Cd^{-1}\rho^d} e^{-\frac{1}{2} r_{ikd}^* [\sum_{t=1}^{\zeta} (x_{td} - m_{ikd})^2]} \, d\, m_{ikd} \\
= \; & (\frac{r_{ikd}^*}{2\pi})^{\frac{\zeta-1}{2}} \cdot (\frac{1}{\zeta})^{\frac{1}{2}} \cdot \frac{\Psi}{2Cd^{-1}\rho^d} \cdot \\
&\{\Phi(\sqrt{\zeta r_{ikd}^*}(m_{ikd}^* - \overline{x}_{\zeta d} + Cd^{-1}\rho^d)) - \\
&\Phi(\sqrt{\zeta r_{ikd}^*}(m_{ikd}^* - \overline{x}_{\zeta d} - Cd^{-1}\rho^d))\}
\end{aligned} \tag{20}$$

where

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{y} e^{-\frac{x^2}{2}} \, dx$$

and

$$\Psi = \exp\{-\frac{1}{2} \zeta r_{ikd}^* [\overline{x_{\zeta d}^2} - (\overline{x}_{\zeta d})^2]\}$$

with $\overline{x_{\zeta d}^2} = \frac{1}{\zeta} \sum_{t=1}^{\zeta} x_{td}^2$ and $\overline{x}_{\zeta d} = \frac{1}{\zeta} \sum_{t=1}^{\zeta} x_{td}$.

4. EXPERIMENTS AND RESULTS

In order to examine the viability of the proposed VBPC algorithm, VBPC is used to perform speaker-independent (SI) recognition of isolated and connected digits, on an isolated Japanese digit database and TIDIGITS English connected digit-string database respectively. In the following experiments, the unknown mismatch is caused by additive Gaussian white noise on the test data. While SI training is performed on clean speech data, in the testing phase, computer-generated Gaussian white noise, with various levels of intensity, is added to the original speech waveform prior to the preprocessing [4]. The degree of mismatch is measured by SNR level (dB) of the contaminated speech, which is calculated averagely over the whole testing set. No knowledge of the above mismatch is explicitly used in testing phase.

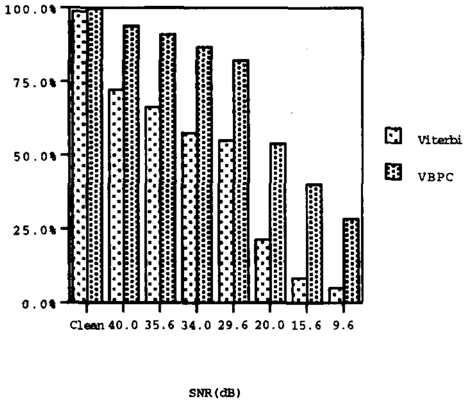

Figure 1. Performance (word accuracy in %) comparison of Viterbi with VBPC at various SNR

Table 1. Recognition accuracy (in %) as a function of neighborhood parameters C and ρ at SNR=34 dB (Viterbi attains 57.5% correct rate at this SNR)

C\ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	57.9	57.9	56.7	57.1	56.3	57.9	63.3	65.8
2	57.5	56.7	56.7	63.3	65.0	71.3	72.9	73.3
3	57.1	56.7	64.2	67.5	68.8	76.3	77.9	79.6
4	56.7	60.4	67.9	69.6	77.1	76.7	76.7	84.6
5	57.9	66.7	69.6	74.2	75.4	74.2	79.2	85.8
6	55.8	67.5	74.2	74.2	73.3	74.2	81.3	85.8
7	58.8	67.9	72.9	73.3	70.0	75.4	80.4	82.1
8	60.4	70.4	71.3	71.3	70.4	76.3	85.8	79.6
9	63.8	72.5	69.2	74.6	70.4	78.3	85.8	74.2
10	66.7	75.4	72.1	73.8	72.1	80.0	85.4	68.3
11	68.3	75.8	70.4	70.8	72.5	78.8	86.7	61.3
12	66.7	75.0	71.3	69.2	70.0	80.4	84.6	59.6
13	67.1	71.3	72.9	68.8	72.5	81.3	84.6	52.9
14	68.8	69.6	71.7	68.8	75.4	82.1	80.8	49.6
15	68.3	69.2	72.1	68.8	74.6	81.7	80.4	41.7
16	67.9	66.3	73.8	68.3	75.4	80.8	79.6	39.6
17	68.8	66.7	74.2	69.6	76.3	80.8	77.9	40.8
18	72.1	66.7	72.9	70.0	77.1	82.5	76.7	34.6
19	71.7	66.3	72.5	70.4	76.7	83.3	75.8	36.3
20	71.7	68.8	71.3	72.1	76.7	85.4	72.9	31.3

4.1. Isolated Digit Recognition

The data is selected from ATR Japanese Speech Database. It contains 0-9 Japanese digit utterances from 60 speakers (half male, half female). The speech was recorded in a quiet environment at sampling rate of 20kHz with 16bit quantisation. Each digit is modeled by a left-to-right 4-state CDHMM without state skipping and each state has 6 Gaussian mixture components with diagonal covariance matrices. Each feature vector consists of 16 LPC-derived cepstral coefficients. For each digit, in total, we have 56 tokens from 46 speakers for SI training, and 24 tokens from other 14 different speakers for SI testing.

Figure 1 compares the averaged recognition accuracy of the VBPC algorithm with that of standard Viterbi algorithm at various SNR levels. The experimental results show that VBPC generally achieves more than 20% recognition rate improvement over that of the conventional Viterbi decoding at various SNR levels. Furthermore, a similar behavior as in minimax approach [4] that the recognition performance tends to be relatively insensitive to the shape of

Table 2. Recognition accuracy (in %) as a function of neighborhood parameters C and ρ on clean data. (Viterbi attains 98.8% correct rate here)

C\ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	98.8	98.3	99.2	99.2	98.8	98.8	99.2	99.2
2	99.2	99.2	99.2	99.2	98.8	98.3	98.8	98.8
3	99.2	98.8	99.2	99.2	98.3	98.3	98.8	97.9
4	99.2	99.2	99.2	99.2	98.3	97.5	97.5	97.5
5	99.2	99.2	98.8	98.3	98.8	97.9	96.7	96.7
6	99.2	98.8	98.8	98.8	98.3	97.1	96.2	92.9
7	99.2	98.8	99.2	98.8	97.9	96.7	96.3	90.8
8	99.2	98.8	98.8	98.8	97.5	96.3	95.4	84.6
9	99.6	98.8	98.3	98.3	97.5	95.4	94.2	79.2
10	99.2	99.2	98.3	97.5	96.7	95.8	93.8	77.9
11	98.8	99.6	98.8	97.5	96.7	95.4	92.9	74.2
12	98.3	99.2	98.3	97.9	97.5	95.4	91.3	64.2
13	98.8	98.8	97.5	96.7	96.3	94.2	89.6	64.6
14	98.3	99.2	97.5	96.7	97.1	95.0	90.0	59.2
15	98.8	98.8	97.5	97.5	96.7	94.6	87.5	53.3
16	98.8	98.8	96.7	97.5	97.1	93.3	86.7	49.6
17	98.8	98.3	97.5	97.1	97.1	94.6	87.9	48.3
18	99.2	97.9	97.1	97.1	97.1	94.6	82.9	45.4
19	99.2	97.5	97.5	97.1	96.3	92.5	80.8	42.9
20	98.8	97.9	97.1	96.3	95.4	90.4	76.7	37.5

uncertainty regions and the performance holds up well under a wide range of SNR values, is also observed in VBPC. As an example, we list the recognition performance as a function of neighborhood parameters C and ρ at SNR=34 dB (mismatched case) and clean data (matched case) in Tables 1 and 2 respectively. Strictly speaking, the performance of VBPC depends on the appropriate choice of ρ and C, which in turn depends on the unknown amount of mismatch. However, the results in Tables 1 and 2 show that considerable improvement (though not optimal) can be obtained in a fairly large range of design parameters (ρ, C), thus suggests that exact knowledge of ρ and C is not crucial.

4.2. Connected Digits Recognition

In contrast with minimax approach, VBPC possesses the intrinsic nature of recursive search, thus VBPC can easily be extended to continuous speech recognition, but with the cost of more computations. VBPC is examined on TIDIGITS corpus which contains utterances from a total of 326 speakers. The SI model for each digit is a 10-state, 10-mixture-per-state CDHMM. The feature vector consists of 12 LPC-derived cepstral coefficients, energy, and their delta features. Because we only consider the uncertainty of Gaussian means in this study, we ignored the contribution of the pre-trained mixture coefficients in VBPC decoding and we found this leads to a better performance in connected digit recognition case. The recognition results of VBPC on TIDIGITS for several SNR levels are listed in Table 3, where **Str** stands for *string correct rate*, **Wd-C** for *word correct rate*, **Wd-A** for *word accuracy*, **Del**, **Sub** and **Ins** for *deletion*, *substitution* and *insertion* error rate respectively.[3] The experimental results show that by using VBPC algorithm, overall recognition performance, say, word(digit) correct rate, is improved about 10% absolutely over that of normal Viterbi decoding.

Table 3. Performance(in %) comparison of Viterbi and VBPC on TIDIGITS corpus

SNR		Str	Wd-C	Wd-A	Del	Sub	Ins
∞	Viterbi	90.0	98.9	97.8	0.3	0.8	1.2
	VBPC	89.5	98.7	97.6	0.4	0.9	1.1
36.8	Viterbi	17.8	67.5	66.4	16.1	16.4	1.1
(dB)	VBPC	37.0	79.4	77.6	8.3	12.3	1.8
27.3	Viterbi	0.2	45.2	43.8	25.2	29.7	1.4
(dB)	VBPC	1.2	53.1	49.4	17.7	29.2	3.7
16.8	Viterbi	0.0	25.1	24.0	45.2	29.7	1.0
(dB)	VBPC	0.0	37.8	32.7	20.8	41.4	5.1

5. DISCUSSION AND CONCLUSION

The above experimental results clearly show that robustness is considerably improved by using VBPC in both isolated word and continuous speech recognition when mismatch exists between test and training conditions. Generally speaking, in the case of less confusable vocabulary where the speech models are distinct enough (ideally no overlap), to use a less *informative* prior distribution such as the uniform distribution we adopted in this study will not cause any problem. Furthermore, it might be beneficial when the mismatch neighborhood described by this prior distribution happens to be consistent with the real mismatch which is the case for additive Gaussian white noise in this study. So the effect of the VBPC decoding is especially pronounced in our experiments. It will be interesting to see how the current VBPC formulation works in other cases of more confusable vocabulary and/or more general unknown mismatches. We are still investigating these issues and will report those results in future.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

[2] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on Speech and Audio Processing*, March 1997.

[3] Q. Huo, H. Jiang, C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *Proc. ICASSP-97*, 1997.

[4] 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.

[5] A. Nadas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 1, pp.326-329, 1985.

[6] R. C. Rose, C.-H. Lee, and B.-H. Juang, "Model compensation for robust ASR," *Proc. IEEE ASR Workshop*, pp.98-100, Snowbird, Utah, 1995.

[7] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. on Speech and Audio Processing*, Vol.4, No.3, pp.190-202, May 1996.

[3] All of these recognition statistics are computed by using HTK.