A MINIMAX SEARCH ALGORITHM FOR CDHMM BASED ROBUST CONTINUOUS SPEECH RECOGNITION

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

[†]Department of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan
[‡] Department of Computer Science & Information Systems, The University of Hong Kong, Pokfulam Road, Hong Kong

ABSTRACT

In this paper, we propose a novel implementation of a minimax decision rule for continuous density hidden Markov model based robust speech recognition. By combining the idea of the minimax decision rule with a normal Viterbi search, we derive a recursive minimax search algorithm, where the minimax decision rule is repetitively applied to determine the partial paths during the search procedure. Because of its intrinsic nature of a recursive search, the proposed method can be easily extended to perform continuos speech recognition. Experimental results on Japanese isolated digits and TIDIGITS, where the mismatch between training and testing conditions is caused by additive white Gaussian noise, show the viability and efficiency of the proposed minimax search algorithm.

1. INTRODUCTION

It is well known now that the mismatches between training and testing conditions will considerably degrade the performance of an automatic speech recognition (ASR) system. How to maintain the recognizer's performance under various mismatches has recently become one of the hottest topics in the area of robust speech recognition. The so-called "compensation/adaptation" approaches [3], which aim at reducing the involved mismatches as much as possible, have formed the mainstream of the current robust speech recognition technology. However, in the past few years, based on robustness theory, some works have been performed to modify the basic decision rule of the ASR system. Instead of directly compensating for the underlying mismatches, the decision rule of the ASR system is designed to be inherently robust to the possible unknown mismatches. This scheme becomes a potential approach for robust ASR because no rigid assumptions about the sources and mechanisms of the mismatches have to be made. Two sets of robust decision rules for ASR, namely, *minimax* decision rule [4, 1] and *Bayesian Predictive Classification* (BPC) rule [1, 2], have been studied. In [4], Merhav and Lee first mentioned the minimax rule in speech recognition community and proposed an implementation for isolated word recognition task. In [1], a so-called *Bayesian minimax* method was proposed to perform a minimax decision under a Bayesian framework. In both of these existing minimax implementations, instead of dynamically searching a desired answer in a structural network representation of all possible hypotheses, decisions are made only from a list of finite candidates. This makes them difficult to be extended to perform continuous speech recognition (CSR) except in an *N-Best rescoring mode.*

In this paper, we combine the idea of the minimax rule with a normal Viterbi search to derive a *recursive minimax search* algorithm for CDHMM (continuous density hidden Markov model) based speech recognition. Because of its intrinsic nature of a recursive search, the approach can be easily extended to perform CSR. A series of experiments are performed on the recognition of isolated digits and TI connected digit strings (TIDIGITS), where the mismatch between training and testing conditions is caused by additive white Gaussian noise (AWGN). The experimental results show that: i) For the isolated digit recognition task, in comparison with the standard Plug-in-MAP method, all three minimax algorithms are able to improve the robustness considerably, while the proposed algorithm performs the best; ii) For connected digit task (TIDIGITS), the proposed minimax search algorithm also achieves a much better performance than that of the conventional Viterbi search algorithm. The increased computational overhead is affordable, at least in this small vocabulary task.

2. TWO PREVIOUS MINIMAX METHODS FOR ROBUST ASR

We model each speech unit W with an N-state CDHMM with parameter vector $\Lambda = (\pi, A, \theta)$, where π is the initial state distribution; $A = \{a_{ij} \mid 1 \leq i, j \leq N\}$ is the transition matrix; and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\cdots,K}$ for each state i. The state observation probability density function (pdf) is assumed to be a mixture of K multivariate Gaussian pdf's with the mixture coefficients ω_{ik}, the D-dimensional mean vectors m_{ik}, and the diagonal precision (inverse covariance) matrices r_{ik}.

In [4], the true parameters of the CDHMM's are assumed to lie within a neighborhood $\eta(\Lambda)$ of the pre-trained models' parameters. Such an uncertainty neighborhood is *parametrically* defined as follows:

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

where constants C $(C > 0)$ and ρ $(0 \leq \rho \leq 1)$ are used to control respectively the possible mismatch *size* and *shape*;

and $\{\pi_i^*, a_{ij}^*, m_{ikd}^*, r_{ik}^*\}$ denote the pre-trained model parameters. Given an observed feature vector sequence X to be recognized, a minimax decision rule is derived in [4] as:

$$\hat{W} = \arg\max_{W} \left[\, p(W) \cdot \max_{\Lambda \in \eta(\Lambda)} \; p(X|\Lambda, W) \,\right] \tag{2}$$

where $\hat{W}$ is the recognition result. In this paper, the above Merhav & Lee's implementation of the minimax decision rule is referred to as minimax1 for convenience.

Another so-called Bayesian minimax rule proposed in [1] works as follows:

$$\hat{W} = \arg\max_{W} p(X \mid \Lambda_{MAP}, W) \tag{3}$$

where

$$\Lambda_{MAP} = \arg\max p(X \mid \Lambda, W) \cdot p(\Lambda \mid \varphi, W) \tag{4}$$

with the prior pdf $p(\Lambda \mid \varphi, W)$ chosen as the best normal approximation to the constrained uniform distribution within the neighborhood $\eta(\Lambda)$ in eq.(1). In this approach, the minimax rule is realized under a Bayesian framework, where the least favorable parameters are obtained by an iterative MAP estimate. This Bayesian minimax method will be denoted as minimax2 thereafter.

As we mentioned before, it is difficult for both the minimax1 and minimax2 to be extended to perform CSR except in an N-Best rescoring mode.

3. MINIMAX SEARCH FOR ROBUST CONTINUOUS SPEECH RECOGNITION

In order to execute the minimax decision rule in robust CSR, we combine the idea of the minimax rule (minimax1) with the normal Viterbi search to derive a recursive minimax search algorithm as follows:

$$\hat{W} = \arg\max_{W} \left[p(W) \cdot \max_{s,l} \max_{\Lambda \,\in\, \eta(\Lambda)} p(X, s, l \,|\, \Lambda, W)\right] \tag{5}$$

where s is the unobserved state sequence and l is the associated sequence of the unobserved mixture component labels corresponding to the observation sequence X.

In our implementation, for every time instant, the least favorable model parameters in the minimax rule are estimated based on each active partial path via only one iteration; then the score of the partial path can be re-computed by using the estimated least favorable parameters accordingly. Based on these re-computed scores, all the active partial paths are propagated in the network in a similar way as in the normal Viterbi search. The recursive minimax search is named as minimax3 in this paper.

Given a test utterance $X = (x_1, x_2, \cdots, x_T)$, CDHMM parameter Λ as well as its corresponding uncertainty neighborhood $\eta(\Lambda)$[1], the recursive search algorithm to *approximately* achieve the minimax3 decision rule in eq.(5) is described as follows:

(1) Initialization ($t = 0$)

$$\alpha_0(i) = \pi_i \quad 1 \leq i \leq N \tag{6}$$

$$\psi_0(i) = 0 \quad 1 \leq i \leq N \tag{7}$$

$$\phi_0(i) = 0 \quad 1 \leq i \leq N \tag{8}$$

where $\alpha_t(i)$ denotes the score of the optimal partial path arriving at state i at the time instant t. The corresponding best partial path is represented by a chain of state points started from $\psi_t(i)$ and a chain of mixture component label points started from $\phi_t(i)$.

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

(2.1) Path-merging in state j:

$$\alpha_t(j) = \max_{1 \leq i \leq N} [\alpha_{t-1}(i) \cdot a_{ij}] \tag{9}$$

$$\psi_t(j) = \arg\max_{1 \leq i \leq N} [\alpha_{t-1}(i) \cdot a_{ij}] \tag{10}$$

$$\phi_t(j) = \arg\max_{1 \leq k \leq K} \; \omega_{jk} \cdot \prod_{d=1}^{D} \sqrt{\frac{r_{jkd}}{2\pi}} e^{-\frac{1}{2} r_{jkd}(x_{td} - \check{m}_{jkd})^2} \tag{11}$$

where

$$\check{m}_{jkd} = \begin{cases} m_{jkd} - Cd^{-1}\rho^d & \text{if } \; x_{td} \leq m_{jkd} - Cd^{-1}\rho^d \\ m_{jkd} & \text{if } \; m_{jkd} - Cd^{-1}\rho^d \leq x_{td} \\ & \qquad \leq m_{jkd} + Cd^{-1}\rho^d \\ m_{jkd} + Cd^{-1}\rho^d & \text{if } \; x_{td} \geq m_{jkd} + Cd^{-1}\rho^d \end{cases} \tag{12}$$

(2.2) Estimate the least favorable parameters Λ^* for all active partial paths:

$$\Lambda^* = \arg\max_{\Lambda \in \eta(\Lambda)} p(x_1, \cdots, x_t, s_{\psi_t(i)}, l_{\phi_t(i)} \mid \Lambda)$$

where $s_{\psi_t(i)}$ and $l_{\phi_t(i)}$ denote respectively the state sequence and the mixture component label sequence corresponding to the active optimal partial path backtracked from the points $\psi_t(i)$ and $\phi_t(i)$. When the neighborhood eq.(1) is adopted, only the mean vectors are adjusted. Thus all the mean vectors $m_{ik}(1 \leq i \leq N, 1 \leq k \leq K)$ of CDHMM are re-estimated as follows:

if (the mixand m_{ik} is included in the partial path $\{\, s_{\psi_t(i)}, \, l_{\phi_t(i)}\}$), **then**

$$\bar{m}_{ikd} = \frac{\sum_{\tau=1}^{t} x_{\tau d}\, \delta(s_{\psi_t(i)}^{(\tau)} - i)\delta(l_{\phi_t(i)}^{(\tau)} - k)}{\sum_{\tau=1}^{t} \delta(s_{\psi_t(i)}^{(\tau)} - i)\delta(l_{\phi_t(i)}^{(\tau)} - k)} \qquad (1 \leq d \leq D) \tag{13}$$

else

$$\bar{m}_{ikd} = m_{ikd} \quad (1 \leq d \leq D) \tag{14}$$

where $s_{\psi_t(i)}^{(\tau)}$ and $l_{\phi_t(i)}^{(\tau)}$ denote respectively the state and mixture component labels corresponding to the time instant τ in the partial path backtracked from the points $\psi_t(i)$ and $\phi_t(i)$.

[1] The neighborhood eq.(1) is still adopted for $\eta(\Lambda)$ here, in which only the uncertainty of mean vectors is taken into account.

Then the least favorable mean vectors are calculated as: (for all $1 \leq i \leq N$, $1 \leq k \leq K$ and $1 \leq d \leq D$)

$$m^*_{ikd} = \begin{cases} m_{ikd} - Cd^{-1}\rho^d & \text{if } \bar{m}_{ikd} \leq m_{ikd} - Cd^{-1}\rho^d \\ \bar{m}_{ikd} & \text{if } m_{ikd} - Cd^{-1}\rho^d \leq \bar{m}_{ikd} \\ & \quad \leq m_{ikd} + Cd^{-1}\rho^d \\ m_{ikd} + Cd^{-1}\rho^d & \text{if } \bar{m}_{ikd} \geq m_{ikd} + Cd^{-1}\rho^d \end{cases} \tag{15}$$

(2.3) Re-score the partial path based on the updated least favorable parameters $\Lambda^* = (\{\pi_i\}, \{a_{ij}\}, \{\omega_{ik}\}, \{m^*_{ikd}\}, \{r_{ikd}\})$:

$$\alpha_t(j) = \pi_{s^{(1)}_{\psi_t(i)}} \cdot \prod_{\tau=1}^{t-1} a_{s^{(\tau)}_{\psi_t(i)} s^{(\tau+1)}_{\psi_t(i)}} \cdot \prod_{\tau=1}^{t} \omega_{s^{(\tau)}_{\psi_t(i)} l^{(\tau)}_{\phi_t(i)}}$$
$$\prod_{d=1}^{D} \sqrt{\frac{r_{s^{(\tau)}_{\psi_t(i)} l^{(\tau)}_{\phi_t(i)} d}}{2\pi}} e^{-\frac{1}{2} r_{s^{(\tau)}_{\psi_t(i)} l^{(\tau)}_{\phi_t(i)} d} (x_{\tau d} - m^*_{s^{(\tau)}_{\psi_t(i)} l^{(\tau)}_{\phi_t(i)} d})^2} \tag{16}$$

(3) Termination

$$s^*_T = \arg\max_i \ \alpha_T(i) \tag{17}$$

(4) Path Backtracking

$$s^*_t = \psi_{t+1}(s^*_{t+1}) \quad t = T-1, T-2, \cdots, 1 \tag{18}$$

The final recognition result $\hat{W}$ can be derived from the optimal path $\{ \ s^*_t \mid t = 1, 2, \cdots, T \ \}$.

Because of its intrinsic nature of recursive search, minimax3 can be easily extended to perform continuous speech recognition. In comparison with the normal Viterbi algorithm, minimax3 needs extra efforts to re-score each active partial path during the search process. However, if the size of the network to be examined is moderate, the increased computational cost is generally affordable.

4. EXPERIMENTS AND RESULTS

In order to examine the viability of the proposed minimax3 algorithm, we present a series of experiments where the minimax3 algorithm is compared with other existing methods. Firstly, in an isolated Japanese digit recognition task, minimax3 is compared with the Plug-in-MAP based Viterbi algorithm, minimax1 and minimax2. As a remark, only a Viterbi version of the minimax2 in [1] is implemented here. Next, in another connected word recognition task on TIDIGITS, minimax3 is compared with the conventional Viterbi search in terms of both the recognition accuracy and computational complexity. In all the experiments, the mismatch between training and testing conditions is caused by adding, at different SNR (signal-to-noise ratio) levels, computer-generated white Gaussian noise (AWGN) into the test data prior to the pre-processing stage. The AWGN is scaled to a fixed level for all utterances in the test set. The degree of mismatch is measured by SNR level (in terms of dB) of the contaminated speech, which is calculated over the whole testing set as follows:

Table 1. Performance (word accuracy in %) comparison of minimax3 with Plug-in-MAP, minimax1 and minimax2 in isolated Japanese digit recognition task when test data are distorted by AWGN.

SNR	Plug-in MAP	minimax1	minimax2	minimax3
∞	98.50	99.58	99.58	99.58
30(dB)	62.08	73.33	71.67	77.50
20(dB)	26.10	57.92	53.33	61.67
10(dB)	5.42	28.33	26.25	33.33

$$\text{SNR} \triangleq 10 \log_{10} \frac{\sum_{i \in \mathcal{S}} \sigma_s(i)}{\sum_{i \in \mathcal{S}} \sigma_n(i)} \tag{19}$$

where $\sigma_s(i)$ denotes the signal variance of the ith speech utterance in test set $\mathcal{S}$, and $\sigma_n(i)$ the variance of noise signal added to the ith utterance. However, no knowledge of the related mismatch is explicitly exploited in testing phase.

4.1. Isolated Digit Recognition: ATR-JPD

In order to compare the performance of minimax3 with other two previous minimax methods (minimax1 and minimax2), we first perform a series of comparative experiments on a speaker-independent (SI) recognition task of isolated Japanese digits on the ATR-JPD database, which is selected from ATR Japanese Speech Database and contains isolated utterances of Japanese 0-9 digits from 60 speakers (half male, half female). The database ATR-JPD is recorded in a quiet environment at a sampling rate of 20kHz with 16bit quantization accuracy. 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.

In Table 1, the averaged recognition accuracy of the minimax3 is compared with that of the standard Plug-in-MAP based Viterbi search algorithm, minimax1, and minimax2 at three SNR levels of 10(dB), 20(dB) and 30(dB). The experimental results clearly show that all three minimax algorithms are able to improve the robustness considerably in comparison with the standard Plug-in-MAP based Viterbi algorithm when the AWGN-caused mismatch exists between the training and testing conditions. We also note that minimax3 significantly outperforms both the minimax1 and minimax2 in the examined SNR levels. This can be explained by the fact that the minimax rule is repetitively applied during the recursive minimax3 search, which warrants to find a better path than both minimax1 and minimax2 in which the minimax rule is only used to re-score the paths found by the normal Viterbi search. We have to note that in Table 1 we only report the optimal performance for all three minimax methods when the hyper-parameters (C, ρ) are manually adjusted within the range:

Table 2. Performance (in %) comparison of minimax3 (mm3) with Plug-in-MAP (PIM) method on TIDIGITS when test data are distorted by AWGN

SNR		Str	Wd-C	Wd-A	Del	Sub	Ins
∞	PIM	88.14	98.44	97.34	0.69	0.87	1.10
	mm3	87.64	98.43	97.20	0.65	0.92	1.22
36.8 (dB)	PIM	17.45	67.37	66.28	16.22	16.40	1.09
	mm3	64.78	95.49	90.0	0.90	3.60	5.49
27.3 (dB)	PIM	0.23	45.25	43.90	25.10	29.70	1.40
	mm3	42.03	87.29	79.03	2.67	10.0	8.26
16.8 (dB)	PIM	0.0	24.89	23.93	45.40	29.70	0.96
	mm3	14.47	66.19	57.52	7.60	26.20	8.70

Table 3. Comparison of the total recognition time (in seconds) of 300 TIDIGITS utterances on a SUN Utra-I workstation between the minimax3 and Plug-in-MAP based Viterbi search

	Viterbi	minimax3
CPU time used (s)	771.94	1538.73

$C \in [1, 10]$ and $\rho \in [0.1, 0.9]$. Besides the optimal performance in Table 1, we also observed in our experiments that minimax1, minimax2 and minimax3 outperform the Plug-in-MAP method for a wide range of (C, ρ). However, we have not found a good method yet to automatically adjust (C, ρ) for the optimal performance in all these minimax methods.

4.2. Connected Word Recognition: TIDIGITS

In order to examine the feasibility of the minimax3 in terms of its computational complexity in a continuous speech recognition task, we also perform a series of comparative experiments of SI connected digits recognition on TIDIGITS English connected digit-string database. Only the part of adult speech data (111 men, 114 women) is used in the experiments. The feature vector consists of 12 LPC-derived cepstral coefficients, energy, and their delta features. Because we are using the delta features, the mean vector m_{ik} consists of static feature in the low dimensions and delta feature in the high dimensions. The uncertainty neighborhood of Λ defined in eq.(1) is slightly modified to take delta feature into account. The SI model for each digit is a 10-state, 10-mixture-per-state CDHMM. All digit HMMs are trained on 8623 utterances from adult training data subset of TIDIGITS. The algorithms are evaluated on 8700 utterances from adult test data subset distorted by various levels of AWGN.

The experimental results[2] in Table 2 show that the minimax3 performs much better than the conventional Viterbi algorithm for the examined SNR levels. As far as the computational complexity is concerned, in minimax3, the increased computation mainly lies in: i) estimating the least favorable parameters as in eqs.(13) and (15); ii)re-scoring the partial path in eq.(16). However, in each search step, only a small portion of the individual partial path need to be re-calculated while the most part of it remains unchanged as in eq.(14). In the experiment, we observed that in this small vocabulary task where the recognition network is not very large, the calculation overhead of the minimax3 is affordable. As an example, we list in Table 3 the total CPU time used by Viterbi and minimax3 full searches (i.e., the beam width is set to infinity) to recognize in total 300 utterances randomly chosen from the test set of TIDIGITS. The CPU time in Table 3 show that the computational complexity is approximately doubled in the minimax3 search in comparison with the normal Viterbi search.

5. DISCUSSIONS AND CONCLUSIONS

From the above experimental results, it is found that given an appropriate uncertainty neighborhood, the robustness of an ASR system can be enhanced by adopting the minimax decision rule. The proposed minimax search algorithm is shown to be effective and efficient for the examined small vocabulary tasks of either isolated words or continuous speech. As future works, we need to develop some methods to automatically determine the hyperparameters (C, ρ) of the uncertainty neighborhood. We should also consider other possibility in uncertainty modeling such as the *distribution uncertainty* in stead of the current practice of the *model parameter uncertainty.*

ACKNOWLEDGMENT

The authors would like to thank Dr. C.-H. Lee, Head of Dialogue Systems Research Department at Bell Laboratories, Murray Hill, USA, for his beneficial comments and discussions on this work.

REFERENCES

[1] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," submitted to *IEEE Trans. on SAP*, August 1997.

[2] H. Jiang, *A Study on Robust Decision Rules in Automatic Speech Recognition*, Ph.D. thesis, the University of Tokyo, June 1998. (available at *http://www.gavo.t.u-tokyo.ac.jp/~jiang/jiang.html*)

[3] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition For Unknown Communication Channels* (Pont-a-Mousson, France), pp.45-54, April 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, Jan. 1993.

[2] 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 rates respectively.