MULTILEVEL DISCRIMINATIVE TRAINING FOR SPELLED WORD RECOGNITION

Luca Rigazio, Jean-Claude Junqua, Michael Galler

Panasonic Technologies Inc. / Speech Technology Laboratory
3888 State Street, Suite 202, Santa Barbara, CA 93105, U.S.A.
email: rigazio, jcj, galler @ research.panasonic.com

ABSTRACT

Discriminative training is effective in enhancing robustness for recognition tasks characterized by high confusion rates. In this paper, we apply discriminative training to different components of a spelled word recognizer to improve recognition accuracy among confusable letters. First we weighted the HMM states to emphasize the letters' discriminant part. The training achieved a 17% decrease in unit (letter) error rate when the search was performed with an unconstrained grammar. Then we designed a new algorithm that relies on discriminative training to adapt the grammar transition probabilities and the language weight. This method uses acoustic information to provide a tight coupling between the acoustic and language models. Experimental results showed the state weighting followed by the adaptation of a bigram language model reduced by 11% the total unit errors and by 12% the unit errors among the E-Set of the English alphabet.

1. INTRODUCTION

Spelled word speech recognition is an interesting topic linked to many applications. The main difficulty is related to confusable letters e.g., E-Set (b,c,d,e,g,p,t,v,z). In this paper, we report results achieved with SmarTspelL$^{\text{TM}}$[1] [1] [2], the Panasonic spelled word recognizer, and focus on optimization of the first-pass decoding. At the acoustic level we attack the problem of confusable letters with discriminative training (DT) techniques. In particular, we estimate the HMM state weights using a minimum string error objective leading to the Discriminative State Weight (DSW) adaptation. By employing DT at the grammar adaptation level we produced a robust estimation of the language model and the language weight. This method has the advantage of taking into account both the training vocabulary and the underlying acoustic behavior, so that the grammar will be tailored to the acoustic models used.

[1] SmarTspelL$^{\text{TM}}$ is a trademark of Panasonic Technologies, Inc.

2. BACKGROUND AND RELATED WORK

DT aims to minimize the *expected error rate* through a discrimination enhancement between the correct and the wrong decoded candidates. The optimization relies on the General Probabilistic Descent (GPD) algorithm, explained in [3] where experiments were performed on isolated E-Set letters and connected digit recognition. In continuous recognition tasks an N-Best competing string objective is generally used for string error minimization [4]. DT techniques have interesting interactions with state-weighting methods employed to highlight the salient section of a word [5]; these two methods can be jointly used to improve the robustness of the acoustic models [6]. In [7] a simple technique is introduced to train the state-weighted models, with experiments on isolated word and connected digit recognition. Regarding the estimation of robust language models on sparse data a well known method is *backing-off* [8] in order to predict the probability of unseen context using the lower order probability. In this paper we apply discriminative methods for continuous letters recognition introducing a multilevel optimization for both the acoustic and the language models.

3. PROBLEM AND DATA DESCRIPTION

SmarTspelL$^{\text{TM}}$ is a multi-pass spelled word recognition system. We are interested in the first-pass decoding, where an HMM alignment is accomplished with a bigram language model. The front-end analysis is performed on a 20 ms window with frame rate of 10 ms by computing 18 parameters: energy, delta-energy, 8 cepstral coefficients and 8 delta cepstral. The acoustic models are whole letter HMMs, with a left-right topology and 3 states for the silence, 12 for the "W", and 8 for all the other letters. The experiments were performed on the spelled name part of OGI [9], recorded on telephone bandwidth of 8 kHz, that we partitioned into three sets, such that each speaker appears in only one set: a training set with 1222 calls, a validation set with 558 calls and a test set with 491 calls. Each set contains several examples of each of the 26 possible letters. We completed the train-

ing set with some utterance from the Macrophone database [10], to cover the examples of infrequent letters. The baseline models were trained with Baum-Welch maximum likelihood re-estimation and the bigram language model was estimated with a backing-off algorithm on the training set transcriptions. Therefore the experiments were completely speaker independent.

4. DISCRIMINATIVE TRAINING

Consider an observation X belonging to one of the classes $C_i, i = 1, 2, \cdots, M$. A classifier defined by the parameters Λ associates with each class C_i a discriminative function $g_i(X;\Lambda)$ and applies the decision rule:

$$C(X) = C_k, \quad k = \arg\max_i g_i(X;\Lambda).$$

In continuous speech recognition the observation X is a vector of short term measurements of the spelled utterance, and discriminative functions commonly employed are the log likelihood (score) of the N-Best strings provided by the decoding algorithm [4]. DT directly optimizes the classifier's expected error rate using gradient descent search. For an X belonging to the class C_i, define the *misclassification measure* with:

$$d(X;\Lambda) = -g_i(X;\Lambda) + \log\left\{\frac{1}{M-1}\sum_{j,j\neq i} e^{g_j(X;\Lambda)\eta}\right\}^{\frac{1}{\eta}},$$

where $\eta > 0$. Then, define the *loss function* with $l(X;\Lambda) = f(d(X;\Lambda))$ where $f()$ is a sigmoid function. An important result of DT theory is that the *expected loss* $E\{l(X,\Lambda)\}$ is a continuous measure asymptotically approaching the expected error rate as $\eta \to 0$. The GPD algorithm minimizes the expected loss through a steepest descent procedure by computing $\Lambda_{t+1} = \Lambda_t - \epsilon_t \nabla l(X;\Lambda_t)$. Therefore GPD asymptotically optimizes the expected error rate as well. Moreover, the corrective term $\nabla l(X;\Lambda_t)$ is related to the discriminative function gradient $\nabla g_i(X,\Lambda)$ [7], that is computed in the next sections for the two training algorithms.

4.1. HMM State Weighting

The basic idea behind state-weighting is that each portion of a speech utterance has different importance for the classification process. The method used to exploit this information consists of the assignment of a weight to each HMM state representing the importance of the emitted frames. Consider the $\mathrm{k^{th}}$ best decoded string $S^k = (s_1^k, s_2^k, \cdots)$ as a sequence of HMMs, let $(q_1^k, q_2^k, \cdots, q_T^k)$ be the state sequence and $\mathcal{L}(q_t^k, x_t|q_{t-1}^k)$ the state-score of observing x_t during the transition $q_{t-1}^k \to q_t^k$. Weighting the HMM states means applying a weight w_q to each state-score. We trained the state-weights with a minimum string error objective by using the weighted state-score associated with the string S^k as a discriminative function:

$$g(X, S^k) = \sum_{t=1}^{T} w_{q_t^k}\, \mathcal{L}(q_t^k, x_t|q_{t-1}^k).$$

We constrained the weights of the model M_i, having N_i states, with $\sum_{q=1}^{N_i} w_q = N_i$ as in [7], and we applied the GPD search in the transformed space $\{\overline{w}_r\}$ defined as in [3] by:

$$w_r = N_i \frac{e^{\overline{w}_r}}{\sum_{q=1}^{N_i} e^{\overline{w}_q}}.$$

The discriminative function derivatives with respect to the transformed state weights are:

$$\frac{\partial g(X, S^k)}{\partial \overline{w}_r} = w_r \left\{ T^k(r;X) + \frac{1}{N_i}\sum_{q=1}^{N_i} w_q T^k(q;X) \right\},$$

where $T^k(r;X) = \sum_{t:q_t^k \equiv r} \mathcal{L}(q_t^k, x_t|q_{t-1}^k)$ is the *cumulative score* for state s along the Viterbi path associated with the string S^k. This term has to be computed for each HMM state for the correct string and for the wrong competing strings with an N-Best search.

4.2. Discriminative Grammar Training

The language weight has a strong influence on the recognition performance because it provides the search algorithm with knowledge about the relative reliability of the language model with respect to the acoustic models. In most current approach the language weight is sub-optimally estimated by means of heuristic tuning on the test set. In this section we derive a GPD-based algorithm to adapt the language weight and we extend it to the grammar transition probabilities. Let us define a stochastic grammar with a graph $\mathcal{G}(\mathcal{N}, \mathcal{A})$ where the nodes $\mathcal{N}$ represent the context and the arcs $\mathcal{A}$ the transitions with which are associated the HMMs. Let $\mathcal{W}(\mathcal{A})$ be the word emitted and $\mathcal{P}(\mathcal{A})$ the probability associated with each transition. Consider the $\mathrm{k^{th}}$ best decoded string $S^k = (s_1^k, s_2^k, \cdots)$ where the $s_i^k \in \mathcal{A}$ are the grammar transitions taken during the utterance decoding. We constrain the transition probabilities p_t with:

$$\sum_{t\in\mathcal{O}(n)} p_t = 1,$$

where $\mathcal{O}(n) \subset \mathcal{A}$ is the set of transitions exiting the node n. Consider now a transition $t \in \mathcal{O}(n)$, we apply the GPD search in the transformed space $\{\overline{p}_t\}$ defined by:

$$p_t = \frac{e^{\overline{p}_t}}{\sum_{r\in\mathcal{O}(n)} e^{\overline{p}_r}}.$$

We assume the discriminative functions are the total score of the strings S^k, expressed by:

$$g(X, S^k) = \sum_{t \in S^k} L \log p_t + A_t,$$

where L is the language weight and A_t is the acoustic score provided by the HMM associated with the transition t. Moreover we assume that the acoustic scores $\{A_t\}$ are independent from the grammar transition probabilities $\{p_t\}$. From the derivatives chain rule we achieve:

$$\frac{\partial g(X, S^k)}{\partial \overline{p}_t} = \sum_{s \in \mathcal{A}} \frac{\partial g(X, S^k)}{\partial p_s} \frac{\partial p_s}{\partial \overline{p}_t}.$$

Consider again a transition $t \in \mathcal{O}(n)$, we can write:

$$\frac{\partial p_s}{\partial \overline{p}_t} = (p_s \delta_{s,t} - p_s p_t) I(s \in \mathcal{O}(n)),$$

where $\delta_{s,t}$ is the Kronecker delta function and I() is the indicator function. From previous assumptions we can derive:

$$\frac{\partial g(X, S^k)}{\partial p_s} = \frac{L}{p_s} \Theta_s^k,$$

where each $\Theta_s^k = \left| r \in S^k : r = s \right|$ counts how many times the transition s has been taken during the string S^k. From previous expressions we achieve:

$$\frac{\partial g(X, S^k)}{\partial \overline{p}_t} = L \left\{ \Theta_t^k - p_t \sum_{r \in \mathcal{O}(n)} \Theta_r^k \right\}.$$

We called this algorithm Constrained Discriminative Grammar Training (C-DGT) since we constrained the grammar probabilities. For the language weight adaptation we perform an unconstrained search so the discriminative function derivative is simply:

$$\frac{\partial g(X, S^k)}{\partial L} = \sum_{t \in S^k} \log p_t.$$

Note that that the two previous terms are also related to the acoustic models since the string S^k is achieved with an alignment of the spoken utterance on the grammar we want to adapt. This method gave the best results when transition probabilities and the language weight were adapted at the same time.

5. EXPERIMENTAL RESULTS

The following results were achieved by re-estimating the baseline acoustic and language models on the training set. To carry out the state weights adaptation the cumulative scores for the correct string were computed with a forced alignment on the utterance labels. For the competing strings, an N-Best decoding was performed with a uniform loop grammar that leaves the search unconstrained and produces a bigger variety of competing strings. The recognition performance achieved with the loop grammar decoding also provides a reliable measure of the acoustic enhancement of the state weighted models. Where possible, we used an exhaustive search useful for smoothing the gradient convergence. During the GPD search we employed a linear decreasing step size applied to the normalized gradient, with maximum iteration as stopping criteria. Fig. 1 shows the string recognition rate achieved with the loop grammar compared with the expected loss over 60 GPD iterations. Although the training is performed with a string error rate objective, for the first-pass decoding we are interested in the unit performance since it can drastically affect the correct string retrieval during post-handling passes. Table 1 shows the improvement in term of unit error reduction rate. The state weight models enhancement, as measured by the loop grammar decoding, is not completely reflected in the decoding performed with the baseline bigram. To deal with this problem we matched the language model and the language weight with the new acoustic models by performing a C-DGT adaptation. The gradient descent terms were computed with an N-Best search on the bigram network for the competing strings, and with a forced alignment on the utter-

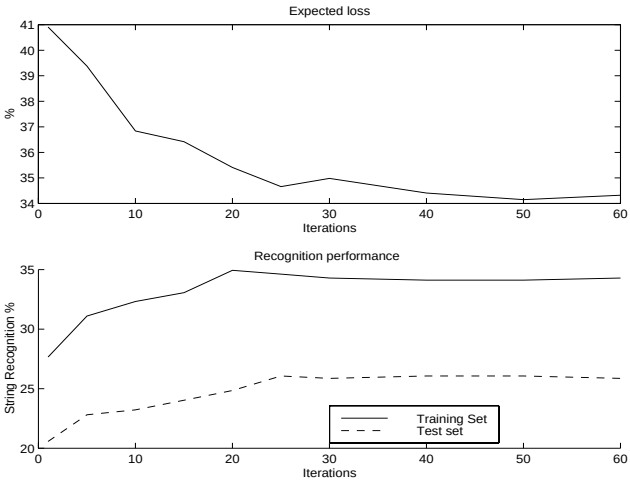

Figure 1: Acoustic model enhancement: string recognition with loop grammar decoding vs. expected loss.

	Train set	Test set
Error reduction	21.1%	17.3%

Table 1: Unit error reduction rate for the DSW adaptation measured with a loop grammar decoding.

	Baseline	DSW	DSW+C-DGT
Insertion rate	1.94%	1.30%	2.03%
Deletion rate	1.02%	1.33%	1.11%
Substitution rate	12.45%	11.88%	10.57%
Unit error rate	15.41%	14.51%	13.71%

Table 2: Unit performance for the DSW and for the DSW followed by the C-DGT adaptations.

ance labels, weighted with the bigram probabilities, for the correct string. Moreover we used two different step sizes, 10 times smaller for the transition probabilities than for the language weight adaptation, to better control the gradient convergence. The adaptation provided an optimal value for the language weight of 14.7 when an exhaustive search performed on the test set gave a value close to 15. Table 2 reports the unit performance achieved with bigram decoding for the baseline models, state-weighted models, and state weighted models with the C-DGT adapted bigram. Notice that the DSW reduced the insertion and the substitution rates at the expense of the deletion rate, and the C-DGT reduced the deletion and the substitution rates at the expense of the insertion rate. By serializing the two adaptations we diminished the negative effects on the insertion and deletion rates and we increased the improvements on the substitution rate. The overall improvements are shown in Table 3. Moreover we obtained a remarkable improvement of 12% for the unit error among the E-Set letters. To illustrate the robustness of this estimate we note that it provide unit performance similar to a backing-off bigram estimated exclusively on the test set.

	DSW	DSW+C-DGT	Total
Error reduction	5.8%	5.5%	11%

Table 3: Unit error reduction rate for the DSW and for the DSW followed by the C-DGT adaptations.

6. CONCLUSIONS

In this paper, we successfully applied discriminative training to the specific task of spelled word recognition. The acoustic model improvements are consistent with previous results in discriminative training for continuous recognition tasks. Moreover, we introduced the discriminative adaptation of language model and language weight and we showed how the multilevel approach profit from improved acoustic models. Note that the C-DGT adaptation was carried out using training data only. A disadvantage of this method is the high computational cost and the need for a corpus of spoken data which well characterizes the language model. By serializing the acoustic models and the grammar adaptation we ended up with high unit performance for the first-pass decoding: 86.3 unit accuracy and 88.3 percent correct. For further improvements it will be interesting to employ discriminative methods at the front-end and in the later search passes.

7. REFERENCES

[1] Junqua J.C. et al, "An N-best strategy, Dynamic Grammars and Selectively Trained Neural Networks for Real-Time Recognition of Continuously Spelled Names Over the Telephone", *ICASSP*, May 1995.

[2] Junqua J.C., "SmarTspelL™ Multipass Recognition System for Name Retrieval over the Telephone", *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, March 1997.

[3] W. Chou, B. H. Juang and C.H. Lee, "Segmental GPD Training of HMM based speech recognizer", *International Conference on Acoustics, Speech, and Signal Processing*, 1992, Vol. 1, pp. 473-476.

[4] W. Chou, C.-H. Lee and B.-H. Juang, "Minimum Error Rate Training based on N-Best string models", *International Conference on Acoustics, Speech and Signal Processing*, 1993, Vol. 2, pp. 652-655.

[5] K.-Y. Su and C.-H. Lee, "Speech recognition using weighted HMM and subspace projection algorithm", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 1, pp. 69-79, 1994.

[6] F. Wolfertstetter and G. Ruske, "Discriminative state-weighting in hidden Markov models", *Proc. Internat. Conf. Spoken Language Processing*, Yokohama, Japan, pp. 219- 222, 1994.

[7] O. W. Kwon, C. K. Un, "Performance of HMM-based speech recognizers with discriminative state-weights", *Speech Communication*, No. 19, pp. 197-205, 1996.

[8] P. Placeway, R. Schwartz, P. Fung and L. Nguyen, "The estimation of powerful language models from small and large corpora", *International Conference on Acoustics, Speech and Signal Processing*, 1993, Vol. 2, pp. 33-36.

[9] R. Cole, K. Roginski and M. Fanty, "A telephone speech database of spelled and spoken names", *ICSLP*, 1992.

[10] J. Bernstein and K. Taussig, J. Godfrey, "Macrophone: an American English telephone speech corpus for the polyphone project", *ICASSP*, 1994.