DISCRIMINATIVE MODEL COMBINATION

Peter Beyerlein

Philips Research Laboratories Aachen, Germany
beyerlei@pfa.research.philips.com

ABSTRACT

Discriminative model combination is a new approach in the field of automatic speech recognition, which aims at an optimal integration of all given (acoustic and language) models into one log-linear posterior probability distribution.
As opposed to the maximum entropy approach, the coefficients of the log-linear combination are optimized on training samples using discriminative methods to obtain an optimal classifier.
Three methods are discussed to find coefficients which minimize the empirical word error rate on given training data:

- the well-known GPD-based minimum error rate training leading to an iterative optimization scheme
- a minimization of the mean distance between the discriminant function of the log-linear posterior probability distribution and an "ideal" discriminant function and
- a minimization of a smoothed error count measure, where the smoothing function is a parabola.

Latter two methods lead to closed-form solutions for the coefficients of the model combination.
Experimental results show that the accuracy of a large vocabulary continuous speech recognition system can be increased by a discriminative model combination, due to a better exploitation of the given acoustic and language models.

1. INTRODUCTION

Given the posterior distribution $\pi(k|x)$, the decision rule that results in a mimimum number of classification errors is the so-called Bayes' decision rule. For a given observation x of unknown class membership, find the class $k(x)$ such that:

$$\forall k' = 1, ..., K; k' \neq k: \quad \log \pi(k|x) - \log \pi(k'|x) \geq 0. \qquad (1)$$

The values $g(x, k, k') = \log(\pi(k|x)/\pi(k'|x))$ in (1) describe the class boundaries and are referred to as discriminant functions [1],[2]. If continuously spoken sentences are recognized the observation is a sequence of feature vectors $x_1^T = (x^1, \ldots, x^T)$, which has to be classified into a word sequence $w_1^S = (w^1, ..., w^S)$. However, the true posterior distribution $\pi(w_1^S|x_1^T)$ is unknown, since it describes the natural speech communication process.
Therefore $\pi(w_1^S|x_1^T)$ has to be approximated by a model distribution
$p(w_1^S|x_1^T)$.

A widely used training criterion for the distribution p is the maximum likelihood criterion. The assumption is that we know the functional form of the probability distribution p, but not the parameters. Using the maximum likelihood criterion the parameters are estimated on training samples. The resulting distribution p is then "plugged in" the Bayes' decision rule: For a given observation x_1^T of unknown class membership, find the class $w_1^S(x_1^T)$ such that:

$$\forall {w'}_1^{S'} \neq w_1^S: \quad \log p(w_1^S|x_1^T) - \log p({w'}_1^{S'}|x_1^T) \geq 0. \qquad (2)$$

Rewriting the discriminant function g

$$\begin{aligned} & g(x_1^T, w_1^S, {w'}_1^{S'}) \\ &= \log p(w_1^S|x_1^T) - \log p({w'}_1^{S'}|x_1^T) \\ &= \log[p(w_1^S)p(x_1^T|w_1^S)] - \log[p({w'}_1^{S'})p(x_1^T|{w'}_1^{S'})], \end{aligned} \qquad (3)$$

we obtain the well-known decomposition of p into a language model probability $p(w_1^S)$ and an acoustic-phonetic likelihood $p(x_1^T|w_1^S)$. Since p typically deviates from the true distribution π, the decision rule (3) will deviate from the Bayes' decision rule, thus leading to a suboptimal classifier. To overcome this limitation discriminative methods can be applied [6],[7]. The goal of discriminative parameter optimization is to be able to correctly discriminate the observations rather than to fit the distributions to the observed data. The most simple example for the discriminative approach is the optimization of the so-called language model factor λ of the discriminant function:

$$\begin{aligned} & g(x_1^T, w_1^S, {w'}_1^{S'}) \\ &= \log[p(w_1^S)^\lambda p(x_1^T|w_1^S)] - \log[p({w'}_1^{S'})^\lambda p(x_1^T|{w'}_1^{S'})]. \end{aligned} \qquad (4)$$

Experiments [5] show that a value λ with $\lambda \neq 1$ gives a minimum word error rate. The deviation from the value $\lambda = 1$ is caused by the deviation of the language model probability $p(w_1^S)$ and the deviation of the likelihood $p(x_1^T|w_1^S)$ from their "true" values.
Following our basic idea we generalize the discriminant function (4). Let us assume that we are given M different acoustic-phonetic and language models $p_j(w_1^S|x_1^T), j = 1, \ldots, M$. These models are log-linearly combined into a distribution of the following form:

$$p^{\Pi}_{\{\Lambda\}}(w_1^S|x_1^T) = e^{\left\{\log C(\Lambda) + \sum_{j=1}^{M} \lambda_j \log p_j(w_1^S|x_1^T)\right\}} \qquad (5)$$

The coefficients $\Lambda = (\lambda_1, ..., \lambda_M)^{tr}$ can be interpreted as weights of the models p_j within the model combination (5). The value $C(\Lambda)$ is a normalization factor. As opposed to the maximum entropy approach [3],[4], which leads to a distribution of the same functional form, the coefficients Λ are optimized with respect to the decision error rate of the discriminant function (6)

$$\begin{aligned} & g(x_1^T, w_1^S, {w'}_1^{S'}) \\ = & \sum_{j=1}^{M} \lambda_j \left(\log p_j(w_1^S|x_1^T) - \log p_j({w'}_1^{S'}|x_1^T) \right) \end{aligned} \quad (6)$$

Note that the discriminant functions in (3) and (4) are special cases of the discriminant function (6). This new approach will be called "Discriminative Model Combination".
In the following it will be shown that a discriminative model combination allows for the integration of any model into a decoder, since the weight λ_j of the model p_j within the combination depends on its ability to provide information for correct classification.

2. DISCRIMINATIVE MODEL COMBINATION

Three methods are discussed to find coefficients Λ, which aim at minimizing the empirical word error rate on given training data:

- The well-known GPD method ('Generalized Probabilistic Descent' [6]) for minimizing the smoothed empirical error rate of the distribution $p^{\Pi}_{\{\Lambda\}}(k|x)$ on training data.
- A minimization of the mean distance between the discriminant function (6) and an "ideal" discriminant function. This method leads to a closed-form solution for Λ.
- A minimization of a smoothed error count measure, where the smoothing function is a parabola.

First the utilized notations are defined:

- Each word sequence w_1^S is interpreted as a class k, each utterance x_1^T is interpreted as an observation x.
- The training data are denoted by $(x_n, k_{nr}), n = 1, ..., N$, $r = 0, ..., K$, where N is the number of acoustic training samples x_n, k_{n0} is the correct class of observation x_n, and $k_{nr}, r = 1, ..., K$ are the competing classes of k_{n0}.
- The value $\mathcal{L}(k_{nr}, k_{n0})$ is the Levenshtein-distance between the rival word sequence k_{nr} and the correct word sequence k_{n0}, i.e. the number of errors contained in the hypothesis k_{nr}

2.1. Minimum Error Rate Training

The Generalized Probabilistic Descent (GPD) algorithm can be applied to minimize the smoothed empirical error rate $L(\Lambda)$:

$$L(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda) \quad (7)$$

on given training data [6]. $\ell(x_n, k_{n0}, \Lambda)$ is a smooth misclassification function of the observation x_n:

$$\begin{aligned} \ell(x_n, k_{n0}, \Lambda)^{-1} = & \quad 1 + \\ & A \cdot \left(\frac{1}{K} \sum_{r=1}^{K} e^{\left\{ -\eta \mathcal{L}(k_{nr}, k_{n0}) \log \frac{p^{\Pi}_{\{\Lambda\}}(k_{n0}|x_n)}{p^{\Pi}_{\{\Lambda\}}(k_{nr}|x_n)} \right\}} \right)^{-\frac{B}{\eta}} \end{aligned} \quad (8)$$

where $A > 0, B > 0, \eta > 0$ have to be adjusted properly. This leads to following iterative scheme to compute the coefficients λ_j with stepsize ε: For $j = 1, \ldots, M$

$$\lambda_j^{(0)} = 1 \quad (9)$$

$$\begin{aligned} & \lambda_j^{(I+1)} = \\ & \lambda_j^{(I)} + \varepsilon \sum_{n=1}^{N} \ell(x_n, k_{n0}, \Lambda^{(I)}) \left(1 - \ell(x_n, k_{n0}, \Lambda^{(I)})\right) \cdot \\ & \cdot \frac{\sum_{r=1}^{K} \mathcal{L}(k_{nr}, k_{n0}) \log \left(\frac{p_j(k_{n0}|x_n)}{p_j(k_{nr}|x_n)} \right) \omega(n,r)^{(I)}}{\sum_{r=1}^{K} \omega(n,r)^{(I)}} \end{aligned} \quad (10)$$

$$\begin{aligned} \omega(n,r)^{(I)} &= \left[p^{\Pi}_{\left\{\Lambda^{(I)}\right\}}(k_{nr}|x_n) \right]^{\eta \mathcal{L}(k_{nr}, k_{n0})} \\ \Lambda^{(I)} &= (\lambda_1^{(I)}, \ldots, \lambda_M^{(I)})^{tr} \end{aligned}$$

From the obtained equation we can see, that the values λ_j are changing from iteration to iteration by a weighted sum of the discriminant function $\log \frac{p_j(k_{n0}|x_n)}{p_j(k_{nr}|x_n)}$.

2.2. Towards A Closed-Form Solution

Since the discriminant function describes the class boundaries it can be argued reasonable to compute values λ_j which minimize the mean squared distance

$$\begin{aligned} D(\Lambda) = & \frac{1}{K \cdot N} \sum_{n=1}^{N} \sum_{r=1}^{K} \\ & \left(\log \frac{p^{\Pi}_{\{\Lambda\}}(k_{n0}|x_n)}{p^{\Pi}_{\{\Lambda\}}(k_{nr}|x_n)} - f(\mathcal{L}(k_{nr}, k_{n0})) \right)^2 \end{aligned} \quad (11)$$

between the discriminant function (6) and the "ideal" discriminant function $f(\mathcal{L}(k_{nr}, k_{n0}))$ on the training data, where $f(.)$ is a squashing function and $\mathcal{L}(k_{nr}, k_{n0})$ is the Levenshtein-distance between the rival word sequence k_{nr} and the correct word sequence k_{n0}.
Optimizing $D(\Lambda)$ by taking the derivatives for the values λ_j we arrive at the following matrix equation for Λ.

$$\Lambda = Q^{-1} P, \quad (12)$$

with

$$Q_{i,j} = \frac{1}{K \cdot N} \sum_{n=1}^{N} \sum_{r=1}^{K} \left\{ \log \frac{p_i(k_{n0}|x_n)}{p_i(k_{nr}|x_n)} \right\} \left\{ \log \frac{p_j(k_{n0}|x_n)}{p_j(k_{nr}|x_n)} \right\},$$

$$(i, j = 1, ..., M),$$

$$P_i = \frac{1}{K \cdot N} \sum_{n=1}^{N} \sum_{r=1}^{K} \left\{ \log \frac{p_i(k_{n0}|x_n)}{p_i(k_{nr}|x_n)} \right\} f(\mathcal{L}(k_{nr}, k_{n0})),$$

$$(i = 1, ..., M). \tag{13}$$

Note, that Q can be interpreted as an $M \times M$ autocorrelation matrix of the discriminant functions of the M given models.
P can be interpreted as correlation vector between the discriminant functions of the M given models and the squashed Levenshtein-distance $f(\mathcal{L}(k_{nr}, k_{n0}))$.
Thus the weight λ_j of a model within the model combination depends upon the correlation of its discriminant function to $f(\mathcal{L}(k_{nr}, k_{n0}))$ and to the discriminant functions of all M models.

2.3. A Closed-Form Solution By Minimization Of A Smoothed Error Count Measure

Define the empirical error rate E:

$$E(\Lambda) = \frac{1}{KN} \sum_{n=1}^{N} \mathcal{L}(\arg\max_{k_{nv}} \left(\log \frac{p^{\Pi}\{\Lambda\}^{(k_{nv}|x_n)}}{p^{\Pi}\{\Lambda\}^{(k_{n0}|x_n)}} \right), k_{n0}),$$

$$= \frac{1}{KN} \sum_{n=1}^{N} \sum_{r=1}^{K} \mathcal{L}(k_{nr}, k_{n0}) \cdot \tag{14}$$

$$\delta(k_{nr}, \arg\max_{k_{nv}} \left(\log \frac{p^{\Pi}\{\Lambda\}^{(k_{nv}|x_n)}}{p^{\Pi}\{\Lambda\}^{(k_{n0}|x_n)}} \right)), \tag{15}$$

This measure equals the classification error rate obtained on the training data by applying the discriminant function (6).
To include all rival hypotheses into the optimization and to get a differentiable cost function, the δ-function in (15) is substituted by a smooth 0-1 funtion $S(x)$

$$E(\Lambda) = \frac{1}{KN} \sum_{n=1}^{N} \sum_{r=1}^{K} \mathcal{L}(k_{nr}, k_{n0}) \cdot S \left(\log \frac{p^{\Pi}\{\Lambda\}^{(k_{nr}|x_n)}}{p^{\Pi}\{\Lambda\}^{(k_{n0}|x_n)}} \right). \tag{16}$$

The contribution of each of the rival hypotheses to the overall smoothed error rate depends on the functional form of S and on the value of the discriminant function for this hypothesis.
To normalize the values of the discriminant function we impose as additional constraint

$$\sum_{j=1}^{M} \lambda_j = 1 \tag{17}$$

A possible choice for the smooth 0-1 function $S(x)$ for $-B < x < A, A > 0, B > 0$ is the following parabola :

$$S(x) = \left(\frac{x + B}{A + B} \right)^2 \tag{18}$$

The values A, B should be chosen such that $-B < \log \frac{p^{\Pi}\{\Lambda\}^{(k_{nr}|x_n)}}{p^{\Pi}\{\Lambda\}^{(k_{n0}|x_n)}} < A$ holds for every pair (n, r) and every normalized Λ.
A parabola has the nice property, that its derivative is a linear function of x, which simplifies the expression for the derivative of the smoothed error rate E and finally allows for a closed form solution for optimal weights λ_j.
Optimizing $E(\Lambda)$, given the normalization constraint (17), by taking the derivatives of the Lagrangian:

$$L(\Lambda) = \frac{1}{KN} \sum_{n=1}^{N} \sum_{r=1}^{K} \mathcal{L}(k_{nr}, k_{n0}) \cdot S \left(\log \frac{p^{\Pi}\{\Lambda\}^{(k_{nr}|x_n)}}{p^{\Pi}\{\Lambda\}^{(k_{n0}|x_n)}} \right) + \alpha \left(\sum_{j=1}^{M} \lambda_j - 1 \right), \tag{19}$$

we arrive analogously to section 2.2 at the following matrix equation for Λ:

$$(\alpha, \Lambda^{tr})^{tr} = B Q'^{-1} P', \text{ with} \tag{20}$$

$$Q'_{0,0} = 0, Q'_{0,j} = 1, Q'_{i,0} = \frac{1}{2}(A + B)^2$$

$$Q'_{i,j} = \frac{1}{K \cdot N} \sum_{n=1}^{N} \sum_{r=1}^{K} \mathcal{L}(k_{nr}, k_{n0}) \left\{ \log \frac{p_i(k_{n0}|x_n)}{p_i(k_{nr}|x_n)} \right\} \left\{ \log \frac{p_j(k_{n0}|x_n)}{p_j(k_{nr}|x_n)} \right\},$$

$$(i, j = 1, ..., M),$$

$$P'_0 = \frac{1}{B}$$

$$P'_i = \frac{1}{K \cdot N} \sum_{n=1}^{N} \sum_{r=1}^{K} \left\{ \log \frac{p_i(k_{n0}|x_n)}{p_i(k_{nr}|x_n)} \right\} \mathcal{L}(k_{nr}, k_{n0}),$$

$$(i = 1, ..., M). \tag{21}$$

3. EXPERIMENTS

Experiments were carried out on the male part of the Wallstreet Journal development and evaluation test sets of 1992 (si_dt_05, si_et_05) with a vocabulary of 5000 words.
Triphone models with Laplacian mixture densities and a grand variance vector, as well as language models were trained on the corresponding WSJ0 training data [8].
A set of hypotheses was created by decoding N-best lists on the development and evaluation data.
Various language and acoustic models were combined into a decoder, using the discriminative model combination approach. The computed combinations were compared with the optimal combinations obtained by an "exhaustive" search. The model combination was optimized on the development set (si_dt_05) and tested on the evaluation set (si_et_05).
Results of experiments using equation (20) are summarized in Table 1. In the first experiment discriminative model combination was applied to compute the optimal language model factor (λ in (4)). The obtained factor was close to the known optimal factor and gave no significant change in the recognition output (Table 1). This validates the approach at least for the automatic computation of an optimal language model factor.
In a second experiment the accuracy of the decoder could be improved by a combination of a word internal triphone system, a crossword triphone system, a bigram, a trigram and a fourgram language model (see Table 1, ww+xw+bg+tg+fg). The improvement was obtained by optimally integrating all 5 models into one decoder, instead of searching for the 'best' of the language models and the 'best' of the acoustic models (see Table 1).

Table 1: Word error rates (in %) using various N-gram language models (bg-bigram, tg-trigram, fg-fourgram), decision tree clustering (ww - word internal triphones, xw - crossword triphones), exh. search - exhaustive optimization of model combination, DMC - automatic optimization by "Discriminative Model Combination"

male	si_dt_05'92	si_et_05'92
ww+bg (exh. search)	9.4	5.4
ww+bg (DMC)	9.5	5.5
xw+bg (exh. search)	8.2	5.2
xw+bg (DMC)	8.1	5.1
xw+tg (exh. search)	7.0	4.0
xw+tg (DMC)	7.0	4.0
xw+fg (exh. search)	6.6	3.5
xw+fg (DMC)	6.6	3.5
ww+xw +bg+tg+fg (DMC)	6.3	3.2

4. CONCLUSION

The proposed discriminative model combination approach aims at an optimal combination of models $p_j(w|x)$ into a distribution of log-linear form.
For the optimization of the model combination a GPD based iterative formulation was derived. In addition a reasonable optimization criterion was found, which leads to a closed-form solution.
Some examples for the application of discriminative model combination were discussed. Integrating one acoustic and one language model (language model factor), several experiments have validated the optimality of the computed coefficients.
Combining 5 acoustic and language models into one decoder leads to an increased accuracy of the decoder, compared to the best pairwise combination of the 5 models. In this case an exhaustive search for the optimal model combination would be prohibitive.
Using a discriminative model combination we are now able to integrate any model into a decoder, since the weight λ_j of the model p_j within the combination depends on its ability to provide information for correct classification.
Finally, the discriminative model combination allows for the search for more complex and more accurate combinations of models of the speech communication process, which will be the subject of further work.

5. REFERENCES

[1] R.O. Duda, P.E. Hart. Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

[2] K. Fukunaga. Introduction to Statistical Pattern Recognition, Second Edition. Academic Press, 1990. pp. 269,270.

[3] J.N. Kapur, H.K. Kesavan. Entropy Optimization Principles with Applications, Academic Press, New York, 1992

[4] T.M. Cover, J.A. Thomas. Information Theory, John Wiley, New York, 1991

[5] K.-F. Lee. The Development of the SPHINX System, Kluwer Academic Publishers, Boston, 1989.

[6] B. H. Juang, W. Chou, C.H. Lee. Statistical and Discriminative Methods for Speech Recognition, in Speech Recognition and Coding - New Advances and Trends, ed. A. J. Rubio Ayuso, J. M. Lopez Soler, Springer-Verlag, Berlin-Heidelberg, 1995

[7] H. Ney. On The Probabilistic Interpretation of Neural Network Classifiers And Discriminative Training Criteria, in press, IEEE Trans. On Pattern Analysis and Machine Intelligence

[8] P. Beyerlein, M. Ullrich, P. Wilcox. Modelling and Decoding of Crossword Context Dependent Phones in the Philips Large Vocabulary Continuous Speech Recognition System, Proc. EUROSPEECH'97, Rhodos, Greece, pp. 1163-1166

[9] P. Beyerlein. Discriminative Model Combination, Theory and Application to Speech Recognition, Philips Research Report No. 1276/97, Philips Research Laboratories Aachen, 1997