A DISCRIMINATIVE TRAINING PROCEDURE BASED ON LANGUAGE MODEL AND DICTIONARY FOR LVCSR

Daniel Willett, Stefan Müller, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: {willett,stm,rigoll}@fb9-ti.uni-duisburg.de

10.21437/Eurospeech.1999-607

ABSTRACT

In today's HMM-based speech recognition systems, the parameters are most commonly estimated according to the Maximum Likelihood criterion. Because of limited training data, however, discriminative objectives provide better parameter estimates with respect to the Maximum A-Posteriori decision used for decoding. The question of which distribution functions to discriminate from which and to what degree is the most crucial when performing discriminative parameter estimation. This is particularly difficult because beside the distribution functions, the recognition procedure is restricted and guided by several other sources of information, such as language model and transition matrices. This paper extends the approach presented in [10] to the case of triphones, refines the theory and estimation of the state-to-state confusion metric and proposes an approximation that allows the application of the approach on context-dependent systems with reasonable computational cost. The evaluation is performed on continuous HMM speech recognition systems for the WSJ0 5k-task. The results prove the practicability of the approach and its extensions.

1. INTRODUCTION

Most commonly, the parameters of HMM speech recognition systems are estimated according to the Maximum Likelihood Estimation (MLE) criterion.

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} p_\lambda(X|W) \tag{1}$$

Training according to this equation can be performed most efficiently with the EM-algorithm which makes it computationally very cheap. The better the assumed family of distributions covers the true distribution and the more training data is available, the closer the pdfs, estimated according to Maximum Likelihood, converge to the true distributions [5]. The more these assumptions are not met, however, discriminative training procedures like the Maximum Mutual Information Estimation (MMIE) can provide more useful parameter estimates [1]. This is due to the fact that the decoding procedure performs a Maximum A-Posteriori decision which the Maximum Likelihood training does not consider.

The objective of the MMIE differs from the MLE by relating the observation's likelihood given the correct transcription to the likelihood of the incorrect ones:

$$\hat{\lambda}_{MMI} = \arg\max_{\lambda} \frac{p_\lambda(X|W)}{p_\lambda(X)} \tag{2}$$

This relation takes into account that the success of the recognition procedure not only depends on a high likelihood of the correct HMM-sequence, but on a low one of possible incorrect sequences just as well.

2. MMIE IN LVCSR

Several publications reported the superiority of the MMIE over the MLE for tasks like isolated word recognition [1] and phonetic recognition [4] where the overall likelihood of an observation according to the statistical models can be determined exactly using

$$p_\lambda(X) = \sum_{all\ \widehat{W}} p_\lambda(X|\widehat{W}) p(\widehat{W}) \tag{3}$$

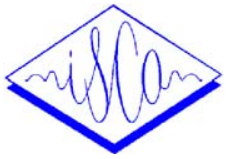
as the number of possible transcriptions is limited.

In continuous speech recognition systems, however, the exact computation of $p_\lambda(X)$, that has to take into account all possible word sequences, is considered to be too complex or it is simply impossible in large vocabulary systems. Hence, methods for the approximation of $p_\lambda(X)$ are required.

Common approaches estimate $p_\lambda(X)$ by considering only the most confusing ("best") sentences which are those sentences W with high posterior probabilities $p_\lambda(W|X)$ which is proportional to $p_\lambda(X|W)P(W)$.

$$p_\lambda(X) = \sum_{all\ \widehat{W}} p_\lambda(X|\widehat{W})P(\widehat{W}) \approx \sum_{best\ \widehat{W}} p_\lambda(X|\widehat{W})P(\widehat{W}) \tag{4}$$

N-best lists [2] and word-lattices [6, 9] have been applied successfully for this approximation. These approaches have two major disadvantages. The one is the computational complexity caused by the need for several n-best or word-lattice recognition procedures on the training data. The other one is the dependence on the confusion measured on the training data that only contains a small fraction of all the possible confusion and that might be specifically distributed misrepresenting the distribution of confusion on unseen test data. This usually results in a remarkable error reduction on the training data while often the effect on test data is very poor.



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A second approach that is computationally much cheaper is to perform the discriminative training on the frame level. Its objective is the optimization of Eq. (5) [1].

$$\hat{\lambda}_{\substack{MMI \\ fb}} = \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{T} \frac{p_{s(i)}(x(i))}{p_M(x(i))} \tag{5}$$

where M represents a general model. This model is usually composed from all models' pdfs according to

$$p_M(x(i)) \approx \sum_{s \in S} p_s(x(i)) P(s) \tag{6}$$

where $P(s)$ represents the HMM states' prior probabilities that can be estimated on the training data.

An algorithm for a discriminative parameter estimation that is based on this frame-based approach has been derived in [10]. It tries to approximate the importance of discrimination of each pair of pdfs by analyzing their contribution in potential word-to-word substitutions that are found in the dictionary. It enhances the frame-based criterion by taking the (possibly wrong) state labeling into account.

$$\hat{\lambda}_{\substack{MMI \\ dict}} = \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{T} \frac{p_{s(i)}(x(i))}{p_M(x(i)|s(i))} \tag{7}$$

where $p_M(x(i)|s(i))$, the general distribution of those x that have been assigned to the state $s(i)$ can further be approximated by

$$p_M(x(i)|s(i)) \approx \sum_{s \in S} p_s(x(i)) P(s|s(i)) \tag{8}$$

with $P(s|s(i))$ representing the probability that an observation assigned to state $s(i)$ has been generated by HMM state s. They indicate the likelihood of mixing up the two HMM states.

Eq. (7) finally results in

$$\hat{\lambda}_{\substack{MMI \\ dict}} \approx \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{T} \frac{p_{s(i)}(x(i))}{\sum\limits_{s \in S, s \neq s(i)}^{S} p_s(x(i)) P(s|s(i))} \tag{9}$$

The probabilities $P(s_1|s_2)$ that represent the importance of discriminating the states s_1 and s_2 can be set up in multiple ways. When approximating them by the states' general priors ($P(s_1|s_2) \approx P(s_1)$), one ends up at the basic frame-based criterion of Eq. (5). [10] proposed to estimate them by analyzing the recognition vocabulary.

The following sections present a refined method of estimating the discrimination weights $P(s_1, s_2)$ for all states $s_1, s_2 \in S$ with $s_1 \neq s_2$, that not only considers similar words found in the dictionary, but also considers sequences of words, state transition probabilities and the language model word likelihoods.

[1] In the equations λ represents the set of all free HMM parameters. Thus, the states' pdfs $p_s(x)$ are defined by this set and should better be written as something like $p_{s\lambda}(x)$. The λ is omitted, however, for the sake of simplicity.

3. ESTIMATING THE STATE-TO-STATE CONFUSION METRIC

The weights $P(s_1|s_2)$ in Eq. (9) represent the dictionary, transition matrices and language model based importance of discriminating the HMM states s_1 and s_2 or, in other words, they represent the danger of mixing up s_1 and s_2 based on all sources of information except of the distribution functions of the acoustical models.

A simple method for setting of approximations for these weights based on the pronunciation dictionary for context-independent recognition systems has been presented in [10]. Its main idea is the analysis of those pairs of words that only differ in one or two models for which the discrimination of the respective acoustical models is particularly important.

	Monophones	Triphones
boot	b u t	b+u b-u+t u-t
bit	b i t	b+i b-i+t i-t
substitutions	1	3

Figure 1. word distances with and without context-dependency

Pairs of words, however, that differ in only one phone in the context-independent case, differ in two or three models in the context-dependent case. Figure 1 illustrates this. Thus, a more profound way for setting up the counts of state-to-state contributions in possible errors has to be established for the case of a context-dependent modeling. Furthermore is it of importance to consider the possible deletions and substitutions by also taking pairs of word sequences into account. Just as much should the transition probabilities be considered as they also have a strong impact on which states are in danger of being mixed up.

The general idea of the refined procedure that we propose with this paper is to set each $P(s_1|s_2)$ proportional to the share of the two states s_1 and s_2 when aligning each word or word sequence (of limited length or likelihood) against each other word or word sequence (of limited length or likelihood). Therefore, the procedure aligns all pairs words or all pairs of word sequences (limited by length or likelihood) against each other with only taking the transition probabilities into account and separately counts the occurrences of overlapping states for each pair of states. The refined procedure that we propose is illustrated in Figure 2.

The outer loops are generating all words, pairs of words or word sequences of limited length. Those words or word sequences that are similar with respect to the (Levenstein) distance of the corresponding model sequences are taken into further consideration.

In the inner part of the procedure the similar words or word sequences are aligned against each other with only taking the transition probabilities into account. The counts of those states that overlap are increased proportional to the words' or word sequences' language model probability and inverse proportional to the distance of the model sequences.

Finally, the state-to-state confusion weights $P(s_1|s_2)$ are set proportional to the accumulated counts C_{s_1,s_2}. The term α_{s_2} can be used to normalize the probabilities to sum

$C_{s_1,s_2} := 0 \quad \forall s_1, s_2 \in S$
x: all words or word sequences (limited by length or likelihood) of the vocabulary {
  y: all words or word sequences (limited by length or likelihood) of the vocabulary {
    d = LevensteinDistance(x, y)
    if $((d \neq 0)$ AND $(d \leq D_{\text{thresh}}))$ {
      $p_x = P_{lm}(x)$; $p_y = P_{lm}(y)$
      l_x = average length (in frames) of x
      align x and y to arrays s_x and s_y of length l_x with each state duration set proportional to the state's average duration
      i: $1 \leq i \leq l_x$ {
        $C_{s_x(i),s_y(i)} += \frac{p_x p_y}{d}$
      }
    }
  }
}
$\forall s_1, s_2 \in S$:

$$P(s_1|s_2) = \begin{cases} \alpha_{s_2} C_{s_1,s_2} & : \quad C_{s_1,s_2} \geq C_{\text{thresh}} \\ 0 & : \quad \text{else} \end{cases}$$

Figure 2. estimation of the language model based state-to-state importance of discrimination in P

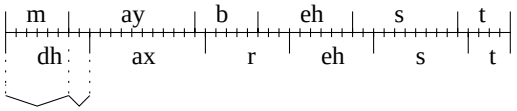
word sequences x = my best y = the rest

phonetically m ay b eh s t dh ax r eh s t

distance $\rightarrow d = 3$ (3 substitutions)

LM-likelihoods $p_x = P_{uni}(my)P_{bi}(best|my)$
$p_y = P_{uni}(the)P_{bi}(rest|the)$

alignment



6 2 time steps

$\rightarrow \quad C_{m,dh} += 6\frac{p_x p_y}{d}$
$C_{ay,dh} += 2\frac{p_x p_y}{d}$
etc.

Figure 3. impact of considering two word sequences on the state counts

up to unity. For the mere optimization of Eq. (9), however, this is not necessary.

For a better understanding of the proposed procedure, Figure 3 illustrates the case of considering the two word sequences 'my best' and 'the rest' in the case of a monophone recognition system of single state Hidden Markov Models.

The threshold D_{thresh} controls the number of words or word sequences that are considered in the inner loop. This way, it directly controls the computational cost and resolution of the whole procedure. The threshold C_{thresh} introduces an approximation that enables to set a large portion of the $P(s_1|s_2)$ to zero, which vastly fastens the training procedure. It should be noted that those components in the dominator of Eq. (9) with $P(s_1|s_2)$ set to zero can be omitted. This enables a speed up of the discriminative training procedure similar to the approach in [7] that limits the discriminative training to the dominant mixture components.

4. TRAINING PROCEDURE

For maximizing discriminative training criteria such as Eq. (9) several approaches can be followed. [3] showed that the EM-algorithm can be extended for the optimization of such rational functions. However, [8] showed that gradient descent optimization procedures accomplish the same task with similar computational cost and success. We chose a gradient descent procedure that optimizes Eq. (9) on a fixed state alignment similar to the RPROP approach outlined in [10].

In order to reduce the computational cost of the training procedure the threshold parameter C_{thresh}, introduced in the previous section, can be applied. In the experiments we found that around 75% of the weights $P(s_1|s_2)$ can be set to zero without a negative impact on the recognition accuracy of the resulting recognition system.

5. EXPERIMENTS AND RESULTS

We used an experimental recognition system for the WSJ0 5k-task for evaluating the proposed approach. The system uses the Limsi-phoneset. It makes use of word-internal triphones of up to 15 Gaussian mixture components each. Decoding is performed with a time-synchronous Viterbi beam search decoder with the standard 5k back-off bigram language model. The system is trained via the EM-algorithm according to the MLE criterion (1).

The recognition accuracy achieved in several experiments is listed in Tab. 1. The baseline performance of the ML-trained system on the Nov'92 development test set of 330 sentences is a word-error rate of 11.3%.

Performing the frame-based discriminative training of Eq. (5), which corresponds to Eq. (9) with the $P(s_1|s_2)$ approximated by $P(s_1)$, results in a measurable but moderate improvement of the word-error rate to 11.0%.

With the weights $P(s_1|s_2)$ estimated according to the procedure given in [10] that only uses the monophone pronunciation dictionary and the triphone state-to-state weights approximated by the corresponding monophone weights, the system achieves a performance of 10.8%.

The preceding column shows the error rate achieved with the weights $P(s_1|s_2)$ estimated with the proposed procedure (Fig. 2) when only considering single word relations and assuming an equal prior probability for each word. The error rate of 10.3% shows that the detailed consideration of the context-dependent models in the approximation of the state-to-state weights is of great importance.

Extending the procedure to the consideration of words

System	word error [%]
baseline ML-traned	11.3
frame-based Eq. 5	11.0
approx. of $P(s_1\|s_2)$ according to [10]	10.8
approx. of $P(s_1\|s_2)$ according to the proposed algorihm (Fig. 2) with only considering single words and using no language model	10.3
as before, but considering single words and word-pairs in the outer loops of Fig. 2 and using no language model	10.3
as before, but with an additional weighting of the words and word-pairs using the bigram language model	9.9
as before, but with the threshold C_{thresh} set very tight ($\approx$ 4 times faster training)	9.9

Table 1. experimental results

and word-pairs (restricted to those with a specific bigram in the back-off language model) and the additional consideration of the words' and word pairs' prior probabilities results in another improvement to an error rate of 9.9%.

These results could also be gained with the threshold C_{thresh} set very strict (with about 75% of the $P(s_1|s_2)$ set to zero) and a dramatically reduced computational cost of the discriminative training procedure.

6. CONCLUSION

The paper has refined the approach for discriminative parameter estimation presented in [10] in several ways. The theoretical foundation has been given that positions the approach in the framework of discriminative training on the frame and on the utterance level. A procedure has been proposed that sets up the state-to-state confusion metric by incorporating the vocabulary and the n-gram language model as well as the HMM states' transition matrices. Approximations have been proposed that reduce the computational cost significantly, so that the discriminative training procedure can even be performed on systems with thousands of (context-dependent) HMM states. The experimental results show the practicability of the approach and its refinements. The best recognition accuracy has been achieved with estimates for the state-to-state importance of discrimination provided by the proposed procedure.

REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer: "Maximum Mutual Information of Hidden Markov Model Parameters for Speech Recognition" Proc. ICASSP'86, pages 49–52.

[2] Yen-Lu Chow: "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition using the N-Best Algorithm" Proc. ICASSP'90, pages 701–704.

[3] P.S. Gopalakrishnan, D. Kanevsky, A. Nádas, D. Nahamoo: "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems" IEEE Transactions on Information Theory, Vol 37, pages 107–113, 1991.

[4] B. Merialdo: "Phonetic Recognition using Hidden Markov Models and Maximum Mutual Information Training" Proc. ICASSP'88, pages 111–114.

[5] A. Nádas: "A Decision Theoretic Formulation of a Training Problem in Speech Recognition, and a Comparison of Training by Conditional versus Unconditional Maximum Likelihood" IEEE Transactions on ASSP, pages 814–817.

[6] Y. Normandin, R. Lacouture, R. Cardin: "MMIE Training for Large Vocabulary Continuous Speech Recognition" Proc. ICSLP'94, pages 1367–1370.

[7] D. Povey, P. C. Woodland: "Frame Discrimination Training of HMMs for Large Vocabulary Speech Recognition" Proc. ICASSP'99, pages 333–336.

[8] R. Schlüter, W. Macherey, S. Kanthak, H. Ney, L. Welling: "Comparison of Optimization Methods for Discriminative Training Criteria" Proc. EUROSPEECH'97, pages 15–18.

[9] V. Valtchev, J.J. Odell, P.C. Woodland, S.J. Young: "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition" Proc. ICASSP'96, pages 605–608.

[10] D. Willett, Ch. Neukirchen, J. Rottland: "Dictionary-Based Discriminative HMM Parameter Estimation for Continuous Speech Recognition Systems" Proc. ICASSP'97, pages 1515–1518.