View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publikationsserver der RWTH Aachen University

A Discriminative Splitting Criterion for Phonetic Decision Trees

Simon Wiesler, Georg Heigold, Markus Nußbaum-Thom, Ralf Schlüter, Hermann Ney

RWTH Aachen University
Chair of Computer Science 6 – Computer Science Department
D-52056 Aachen, Germany
{wiesler, heigold, nussbaum, schlueter, ney}@cs.rwth-aachen.de

Abstract

Phonetic decision trees are a key concept in acoustic modeling for large vocabulary continuous speech recognition. Although discriminative training has become a major line of research in speech recognition and all state-of-the-art acoustic models are trained discriminatively, the conventional phonetic decision tree approach still relies on the maximum likelihood principle. In this paper we develop a splitting criterion based on the minimization of the classification error. An improvement of more than 10% relative over a discriminatively trained baseline system on the Wall Street Journal corpus suggests that the proposed approach is promising.

Index Terms: discriminative training, phonetic decision trees, state tying, new paradigms

1. Introduction

A key issue in all pattern recognition systems is to find a good balance of the model complexity, which on the one hand needs to be high enough to distinguish between the classes and on the other hand to be limited to avoid overfitting. In large vocabulary continuous speech recognition (LVCSR), context dependent (CD) phones, typically tri- or pentaphones, are used as the basic modeling unit. However, modeling all CD phones explicitly would result in an extremely high model complexity. This problem becomes especially severe when across word context is used and the majority of the CD phones is not seen in training at all. A solution to this problem is to tie the parameters of acoustically similar CD phones. Phonetic decision trees [1] are by far the most popular tool for finding such a tying and hence are a key component of all state-of-the-art speech recognizers.

A phonetic decision tree is a binary tree, of which every node corresponds to a phonetic question which implies a partition into a new left and right node. The leaves of the tree correspond to sets of CD phones whose parameters are tied. The construction of the phonetic decision tree is determined by three elements: the set of phonetic questions, a splitting criterion, and a stopping criterion. Some research has been done on the definition of appropriate questions (e.g. [2]) and on the automatic generation of questions [3, 4]. Different stopping criteria have been proposed by [5, 6]. This paper addresses the second issue, the splitting criterion.

In the original approach [1] the training data corresponding to one node is described by a Gaussian distribution and the splitting criterion is defined as the increase in likelihood of the data. In state-of-the-art systems maximum likelihood (ML) acoustic models are only used as an initialization for the discriminative training as for example the minimum phone error (MPE) training [7]. Nevertheless, we only know of a single publication [8] where a discriminative criterion is used for constructing the phonetic decision tree. Furthermore, in all publications except of [6], the performance of the phonetic decision tree is evaluated with suboptimal ML trained recognizers. Even in [8], the acoustic model is trained according to the ML criterion. This inconsistency is probably one of the reasons why their approach performs worse than the conventional splitting criterion. In this paper we propose two discriminative splitting criteria. One of them leads to improvements over a discriminatively trained baseline system.

The remaining paper is organized as follows. In Section 2, the conventional splitting criterion of [1] is briefly reviewed. In Section 3 we propose to base a splitting criterion on the Gaussian classification error. In Section 4 we derive two splitting criteria. The first aims at minimizing the triphone classification error, the other includes a heuristic for minimizing the word classification error. In Section 5, experimental results are presented. The paper concludes with a discussion of the results and an outlook.

2. Maximum likelihood splitting criterion

In order to achieve accurate acoustic models, the states of hidden Markov model (HMM) based speech recognizers include phonetic context. For simplification we only consider triphone states here although the presented concepts are not restricted to that. Let $\mathcal{T}$ denote the set of all triphone states and $\mathcal{X}$ the acoustic vector space. A *state tying* $\mathcal{S}$ is a partition of $\mathcal{T}$ such that the emission probabilities of all triphone states of an $s \in \mathcal{S}$ have the same parameters.

The conventional phonetic decision tree aims at maximizing the log-likelihood of the training data, where the data of one tied state is assumed to be Gauss-distributed with diagonal covariance matrix. Given a set of labelled acoustic vectors $(x_n, \tau_n)_{n=1,\dots N} \subset \mathcal{X} \times \mathcal{T}$ the optimal maximum likelihood state tying is determined by

$$\hat{\mathcal{S}} = \operatorname*{argmax}_{\mathcal{S}} \max_{\theta} \sum_{n=1}^{N} \log p_\theta(x_n|\tau_n). \tag{1}$$

Finding a globally optimal state tying is intractable, because of the huge combinatorial complexity of this problem. In phonetic decision trees this criterion is maximized only locally. In every node F of the tree a set of predefined phonetic questions is posed which define a partion into a left node L and right node R. The algorithm selects the split which gives the biggest improvement in likelihood. This score has a closed form solution and can be calculated from the sufficient statistics of the triphones.

The main advantages of phonetic decision trees are on the one hand their efficiency and simplicity and on the other hand their generalization ability to unseen triphones. However, the key

concept of the conventional phonetic decision tree is the ML principle, even if ML is not used for the training of the emission probabilities. In the following two sections we propose two splitting criteria which are motivated by the classification error.

3. Towards a classification error based splitting criterion

Many discriminative training criteria as the well known maximum mutual information (MMI) criterion can be motivated by being a smooth upper bound to the Bayes classification error [9, 10]. Ideally, a discriminative splitting criterion should minimize the global classification error. The problem of any discriminative approach for the construction of phonetic decision trees is that conventional discriminative criteria consider all classes whereas a key property of phonetic decision trees is their locality. We circumvent this problem by only considering the classification error of the two tied states with the same father node.

The basic assumption of the conventional splitting criterion is that the data corresponding to one node is normally distributed. This assumption allows for the efficient calculation of the increase in likelihood obtained by a split of a node F into the nodes L and R. Here, we further assume that the covariance matrix of the two nodes is identical. Supposed the true distribution equals the model distribution, the error of the binary classification problem can be calculated analytically [11]:

$$\begin{aligned} p(\text{error}_{\mathcal{X}\to\{L,R\}}) = &\frac{p(L|F)}{\sqrt{2\pi}} \int_{\frac{u+t}{v}}^{\infty} \exp(-x^2/2)\,\mathrm{d}x \\ &+\frac{p(R|F)}{\sqrt{2\pi}} \int_{\frac{u-t}{v}}^{\infty} \exp(-x^2/2)\,\mathrm{d}x, \end{aligned} \tag{2}$$

where

$$v^2 = 2 * u = (\mu_L - \mu_R)^t \Sigma^{-1} (\mu_L - \mu_R)$$

is the *Mahalanobis distance* between the mean of the left node μ_L and the right node μ_R and t is the logarithm of the ratio of the priors. The integrals can be calculated efficiently with numerical implementations of the error function. In our implementation we do not use the prior information t since it is not used during recognition, too. Hence, the two integrals coincide.

Since the model assumptions made here are not met in practice, the classification error could be reduced by a discriminative training of the parameters of the classifier. Such an approach requires to process all features for each question that is posed. Its computational costs are therefore very high. Nevertheless, we did some experiments in which we retrained the parameters of the two nodes discriminatively, but the performance of such a decision tree was always worse than that of a decision tree based on Equation (2). A reason might be that in this approach the restriction of considering only the two neighboring nodes is too strong.

The classification error is a measure for the separability of the tied states. However, the separability alone would not lead to a useful splitting criterion, because the goal of the acoustic model is to classify the triphones instead of simply the tied states. This conceptual problem can be illustrated with the trivial state tying that ties all states. Naturally, the single tied state would be classified correctly in all cases, but provides no information about the triphone. Hence, a discriminative splitting criterion needs to reflect the separability of the states as well as their informativeness. In our approach we define a splitting criterion as the sum of the classification error and a term reflecting the informativeness of the tree. Two alternative definitions of such a term are derived in the next section.

4. The gain of a split

4.1. Triphone gain

In order to measure the informativeness of the state tying, we formally consider the classification problem from a tied state $s \in \mathcal{S}$ to a triphone state $\tau \in \mathcal{T}$. The equivocation of $\mathcal{T}$ given $\mathcal{S}$

$$H(\mathcal{T}|\mathcal{S}) = -\sum_{s\in\mathcal{S},\tau\in\mathcal{T}} p(s,\tau)\log p(\tau|s)$$

is an upper bound to the error of this classifier [9]. It can further be simplified to

$$\begin{aligned} p(\text{error}_{\mathcal{S}\to\mathcal{T}}) \le H(\mathcal{T}|\mathcal{S}) = &\quad H(\mathcal{S}|\mathcal{T}) - H(\mathcal{S}) + H(\mathcal{T}) \\ = &\quad 0 - H(\mathcal{S}) + \text{const}(\mathcal{S}). \end{aligned}$$

Here $H(\mathcal{S})$ denotes the entropy of $\mathcal{S}$. The change in entropy caused by a split is

$$\Delta H = -\sum_{S=L,R} p(S)\log p(S) + p(F)\log p(F). \tag{3}$$

The term ΔH can be calculated from the triphone state probabilities $p(\tau)$:

$$p(S) = \sum_{\tau\in S} p(\tau), \quad S = F, L, R \tag{4}$$

Following this approach, a discriminative splitting criterion should minimize the score

$$\mathcal{G}_{\text{triphone}}(F,L,R) = p(\text{error}_{\mathcal{X}\to\{L,R\}}) - \Delta H. \tag{5}$$

The main idea of the splitting criterion (5) is to consider the error of the *triphone-classifier*. The term ΔH is maximized by a tree with equal prior probabilities. Intuitively, it enforces that the number of triphones in one node does not deviate too much from the number of triphones in another node. Experimental results of this approach are presented in Section 5. They show that the proposed splitting criterion performs similar to the conventional splitting criterion, which uses exactly the same statistics. In the next subsection, we propose a splitting criterion that better reflects the goal of recognizing *words*.

4.2. Word gain

The goal of this subsection is to replace the entropy term ΔH by a word gain. The entropy term was motivated by the error of the classifier from $\mathcal{S}$ to $\mathcal{T}$. The idea of our proposed word gain is to consider the error of the classifier from a single tied state s to a word in the recognition vocabulary $\mathcal{W}$. A simple heuristic better reflects this goal. Instead of balancing the number of triphones per node, we balance the number of words containing the triphone. Formally, we define

$$p_{\text{LM}}(S) = \sum_{w\in\mathcal{W}:S\in w} p_{\text{LM}}(w), \quad S = F, L, R,$$

where $p_{\text{LM}}(w)$ is the unigram language model probability of w. The set of words containing the tied state S can be determined

Table 1: Statistics for the Wall Street Journal 5k corpus (WSJ0)

	WSJ0		
	training	**dev**	**eval**
amount of acoustic data [h]	14.77	0.46	0.4
# sentences	7240	410	330
# words	130976	6784	5353

from a pronunciation lexicon, where for simplicity unique pronunciations are assumed. The entropy term ΔH is modified to

$$\Delta H_{\mathrm{LM}} = -\sum_{S=L,R} p_{\mathrm{LM}}(S)\log p_{\mathrm{LM}}(S) + p_{\mathrm{LM}}(F)\log p_{\mathrm{LM}}(F).$$

Combining this term with the Gaussian classification error leads to our proposed splitting criterion analog to Equation (5)

$$\mathcal{G}_{\mathrm{word}}(F,L,R) = p(\mathrm{error}) - \Delta H_{\mathrm{LM}}. \qquad (6)$$

For the calculation of $p_{\mathrm{LM}}(S)$ the set of all words containing S is needed. The sets corresponding to the triphones can be calculated beforehand. During the construction of the tree the sets of one node just have to be combined. The computational costs for this splitting criterion are similar to that of the conventional criterion.

The term H_{LM} has some connection to the idea of a paper about using phone sequences of variable length as features in a log-linear model [12]. Like in our approach, the importance of phone sequences is measured by statistics of the language model. But the approach of [12] is not in the context of state tying.

An advantage of the word splitting criterion is that the triphones are not treated equally as in the previous subsection, but according to their importance for recognizing words. Essentially, triphones in short words get more weight than triphones in long words. Furthermore, the criterion makes use of the very reliable language model statistics during acoustic model training. In the next section, we present experimental results that show the effectiveness of this approach. Nevertheless, a limitation of this method is that it is yet not applicable for across word models, because the pronunciations are just taken from the pronunciation lexicon which does not include context. Including across word information into the criterion for example via bi- or trigram statistics remains an open task.

5. Experimental results

All speech recognition experiments were performed on the WSJ0 corpus with a vocabulary of 5k words. The training corpus consists of 15 hours and the evaluation corpus of 0.4 hours of read speech (see Table 1). The amount of training and test data of this corpus is quite small, but we decided to use this corpus, because changing the decision tree involves a re-training of the acoustic model including the discriminative training. Since this is very costly, WSJ0 is a good choice for this initial study. Since the official WSJ0 corpus does not provide a development set, 410 sentences were extracted from ten new speakers of the North American Business task and used as a development set. All recognition systems were tuned on this development corpus and then applied to the evaluation corpus. The task has a closed vocabulary, that means all words in the evaluation corpus are known.

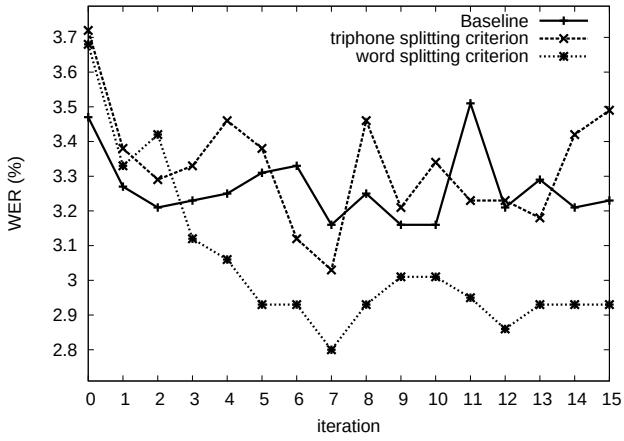

Figure 1: WER for the different iterations of the MPE training on the WSJ0 evaluation corpus

In all systems an acoustic front end consisting of 16 Mel-frequency cepstral coefficients (MFCC) and a voicedness feature was used. Features from nine consecutive frames were concatenated and the resulting vector was projected to a 33 dimensional vector by means of a linear discriminant analysis (LDA). In all experiments a global decision tree was used, but the order of the questions was restricted. First, questions regarding the central phoneme were posed, second, questions regarding the state index of the HMM, and finally questions regarding the triphone context. The growing of the tree was stopped when the number of leaves reached 1500. The systems use within words models only, because the gains of across word modeling on read speech are only small and the small amount of training data does not allow for the use of many tied states with across word context.

The emission probabilities are modelled by Gaussian mixture distributions with a total of about 230 000 densities, all sharing a single diagonal covariance matrix. Since the LDA and the phonetic decision tree mutually depend on each other, both have to be updated iteratively. For initialization, a phonetic decision tree with MFCC features and its first and second derivatives can be used. The performance of the conventional decision tree does not increase after already two iterations. For the discriminative splitting criteria we observed a slight improvement by increasing the number of iterations.

The initial acoustic models were trained according to the ML criterion. Afterwards an MPE training with a margin term as described in [13] was performed. A trigram language model has been used in all recognitions. The ML baseline system achieves a word error rate (WER) of 3.5% (see Table 2). This is much better than the result reported in [14] (4.9 %) and slightly better then the result in [15] (3.9 %), where no VTLN is used. The MPE objective function was optimized with the Rprop algorithm. All parameters of the discriminative training including the initial step size, the I-smoothing-parameter and the choice of the best iteration were tuned on the development corpus. The MPE training improves the baseline system to 3.3 % WER. The absolute improvement of 0.2 % is quite small, but in accordance to the relative gains by discriminative training of e.g. [15] with MMI.

Analogous to the baseline system, we built two systems with phonetic decision trees based on the splitting criteria of Section 4. The error rates on the evaluation corpus for all MPE iterations are depicted in Figure 1. It is noticeable that for all

Table 2: WER on the WSJ0 evaluation corpus of the models corresponding to the best iteration on the development data

	ML	MPE
Baseline	3.5	3.3
triphone splitting criterion	3.7	3.4
word splitting criterion	3.7	2.9

systems the error rate is strongly fluctuating. This behavior may be an indication for a too large step size of the optimization algorithm, but it was not observed on the development corpus. In practice the best iteration has to be determined on the development data. The WER of these models on the evaluation corpus can be found in Table 2. With ML training only, both proposed discriminative splitting criteria perform slightly worse than the conventional splitting criterion. This is not satisfying, because the idea to base the state tying on the classification error should in principle improve a ML trained system, too. The WER of the discriminative state splitting criterion undergoes that of the baseline for some iterations, but for the decisive iterations, the WER of the baseline is slightly better. The result of the discriminatively trained system with the word splitting criterion is 2.9% which clearly outperforms the discriminatively trained baseline system. This is a remarkable improvement of more than ten percent relative. The improvement over the ML baseline system is more than fifteen percent relative, which is more than could be expected of a discriminative training alone.

6. Discussion and Outlook

Considering the big theoretical efforts to improve discriminative training, it seems very promising to optimize the phonetic decision tree in the context of discriminative training, where it has previously just been accepted as given. Also, the computational costs of discriminative training range from hours to days, whereas just a few minutes are spent on the calculation of phonetic decision trees, although the state tying is a crucial aspect of the recognition system.

In this paper, we proposed two alternative splitting criteria for phonetic decision trees. Their goal is a better capturing of the discriminative information among the states. One of the criteria is based on triphone classification, the other on word classification. We could achieve a remarkable improvement over the discriminatively trained baseline system with the word splitting criterion. Although this criterion originally is introduced for within word models, the results show that there is room for improvement of the phonetic decision tree in the context of discriminative training which is not exploited by state-of-the-art systems.

A number of aspects make the definition of a discriminative splitting criterion difficult. First of all, minimizing the tied state classification error alone does not lead to a meaningful criterion. We solved this problem by adding a balancing term, the triphone respectively word gain.

Ideally, the state tying and the tied parameters should be optimized simultaneously with respect to a discriminative criterion. In practice, this is not tractable or requires very strong approximations. In our approach, we first used the ML parameters to minimize the classification error with respect to the state tying. In a second step, the state tying was kept fix and the parameters were optimized with respect to a discriminative criterion.

Finally, the locality of phonetic decision trees contradicts the idea of discriminative training to optimize all classes simultaneously. In our approach we tackled this problem by just considering the two competing classes of one node, which allows for the efficient calculation of the error probability. Certainly, a splitting criterion which is based rigorously on the global classification error would be desirable.

Future work includes the generalization of our proposed word splitting criterion to across word models and the evaluation of both criteria on larger tasks.

7. Acknowledgements

This work was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation

8. References

[1] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Workshop on Human Language Technology*, 1994.

[2] W. Reichl and W. Chou, "A unified approach of incorporating general features in decision tree based acoustic modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999.

[3] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[4] P. Chou, "Optimal partitioning for classification and regression trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 340–354, 1991.

[5] H. Nock, M. Gales, and S. Young, "A comparative study of methods for phonetic decision-tree state clustering," in *European Conference on Speech Communication and Technology*, 1997.

[6] X. Liu and M. Gales, "Model complexity control and compression using discriminative growth functions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

[7] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[8] S. Gao and C. Lee, "A discriminative decision tree learning approach to acoustic modeling," in *European Conference on Speech Communication and Technology*, 2003.

[9] J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, 1970.

[10] H. Ney, "On the relationship between classification error bounds and training criteria in statistical pattern recognition," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2003.

[11] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[12] G. Zweig and P. Nguyen, "Maximum Mutual Information Multiphone Units in Direct Modeling," in *Interspeech*, 2009.

[13] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *International Conference on Machine learning*, 2008.

[14] Z.-J. Yan, B. Zhu, Y. Hu, and R.-H. Wang, "Minimum word classification error training of HMMs for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[15] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Interspeech*, 2005.