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

A COMBINED MAXIMUM MUTUAL INFORMATION AND MAXIMUM LIKELIHOOD APPROACH FOR MIXTURE DENSITY SPLITTING

Ralf Schlüter, Wolfgang Macherey, Boris Müller and Hermann Ney

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology
Ahornstraße 55, 52056 Aachen, Germany
schlueter@informatik.rwth-aachen.de

10.21437/Eurospeech.1999-309

ABSTRACT

In this work a method for splitting continuous mixture density hidden Markov models (HMM) is presented. The approach combines a model evaluation measure based on the Maximum Mutual Information (MMI) criterion with subsequent standard Maximum Likelihood (ML) training of the HMM parameters. Experiments were performed on the *SieTill* corpus for telephone line recorded German continuous digit strings. The proposed splitting approach performed better than discriminative training with conventional splitting and as good as discriminative training after the new splitting approach.

1. INTRODUCTION

An important observation is that the improvements obtained by discriminative training methods in comparison with conventional Maximum Likelihood (ML) training are especially high for low model complexity. Therefore discriminative training criteria might be a good candidate to give an estimation of the ability of an acoustic model to describe the data.

The standard approach to acoustic modeling in speech recognition uses Hidden Markov Models (HMM) in combination with continuous mixture densities. When using mixture densities, a crucial point is the choice of the model complexity, i.e. the determination of the number of densities to be assigned to each mixture model. The usual splitting methods try to double the number of densities iteratively as far as enough observations are assigned to a density. On the one hand one would expect to increase the number of densities for a given mixture with the heterogeneity of the according distribution of the acoustic data. On the other hand this number is clearly limited by the amount of data available for a given task. In order to take this into account, likelihood threshholds could be used to limit the number of densities to be splitted.

In [1] a discriminative measure for model complexity evaluation was introduced. For each mixture model this measure was used to choose the optimal model from ML trained models with differing number of densities. It could be shown that nearly equal performance could be obtained with significantly lower numbers of parameters. Furthermore, in [4] a mixture splitting algorithm fully based on the Maximum Mutual Information (MMI) criterion is introduced. There, a discriminative measure derived from the MMI criterion is used to choose those densities, which are to be splitted up. This was then combined with MMI training of the according models. For connected digit recognition, significant improvements in sentence error rate were observed with this approach when compared to both ML and subsequently MMI trained models of even higher complexity. A similar approach [8] was chosen for large vocabulary speech recognition with similar success, although the approach was not pursued up to optimal model complexities.

Based on the observation, that the advantages of discriminative training criteria diminish with increasing model quality, in this paper we present a method, which combines the use of an MMI based measure to evaluate densities for splitting with subsequent ML training of the according upsplitted models. In our approach, the derivatives of the MMI criterion with respect to the density weights are sorted in order to obtain their median and only those densities with derivatives higher than the according median will be splitted. Experiments on the *SieTill* corpus for telephone line recorded German continuous digit strings show that the combined MMI/ML splitting approach gives better results than conventional splitting with discriminative training and equal results as a subsequent discriminative training after the new splitting approach.

2. DISCRIMINATIVE TRAINING

Before introducing the discriminative splitting criterion proposed here, we will revisit the unified approach for discriminative training proposed in [7]. Let $X_r = x_{r1}, x_{r2}, ..., x_{rT_r}$ and $W_r = w_{r1}, w_{r2}, ..., w_{rN_r}$ denote the sequences of acoustic observation vectors and corresponding spoken words of utterances $r = 1...R$ of the training data. The acoustic emission probability for a word sequence W shall be denoted by $p_\lambda(X_r|W)$, with λ the set of all parameters of the acoustic model. For the following, the language model probabilities $p(W)$ for word sequences W are supposed to be given. For discriminative training we further define the set of alternative word sequences $\mathcal{M}_r$, which are considered for discrimination in utterance r, a smoothing function f and a smoothing exponent α. Using these definitions, we define the following unified discriminative training criterion,

$$\mathcal{F}(\lambda;\alpha, f, \{\mathcal{M}\}) = \sum_{r=1}^{R} f\left(\log \frac{p^\alpha(W_r)p_\lambda^\alpha(X_r|W_r)}{\sum\limits_{W\in\mathcal{M}_r} p^\alpha(W)p_\lambda^\alpha(X_r|W)}\right),$$

Depending on α, f and $\{\mathcal{M}\}$, criteria included in this approach are the Maximum Likelihood (ML), the Maximum Mutual Information (MMI) and the Minimum Classification Error (MCE) criterion. The according choices of $(\alpha, f, \{\mathcal{M}\})$ are summarized in Table 2. Since we want the criterion to be maximized in any case, we take the *negative* sigmoid function for the MCE case. Corrective training [5] is an approximation to the MMI criterion, where only the best recognized word sequence is considered for discrimination. Similarly, using the best scored but *incorrectly* recognized word sequence only we call falsifying training (FT) as a limiting case to MCE training for $\alpha \rightarrow \infty$. In all practical cases the set of alternative word sequences $\mathcal{M}_r$ is obtained by a recognition pass.

An optimization of the unified discriminative training criterion leads to a simultaneous maximization of the emission probabilities of the spoken word sequences and minimization of the weighted sums over the emission probabilities of each allowed alternative word sequence given the acoustic observation sequence for each training utterance. In other words, discriminative training optimizes class separability according to the choices of alternative word sequences and the language model.

Table 1: Settings of the set of alternative words for discrimination M_r and smoothing function f and exponent α for several criteria.

criterion	smoothing function $f(z)$	word sequences included in $\mathcal{M}_r$	exponent α
ML	identity	–	obsolete
MMI	identity	all	1
CT		best recogn.	∞
MCE	$-\frac{1}{1+e^{2\varrho z}}$	all *without* W_r	*free*
FT		best recogn. $\neq W_r$	∞

2.1. Parameter Optimization

Before introducing reestimation equations derived from the unified criterion, we will define discriminative averages, which make use of the following definitions. The mixture density for an acoustic observation vector x given an HMM state s shall be denoted by $p(x|s,\lambda_s)$. The according parameters λ_s of the mixture density identify the weights c_{sl} and parameters λ_{sl} of densities l of the mixture. For the case of maximum approximation considered here we further introduce density probabilities being equal to 1 for the best density of a mixture and zero otherwise:

$$\eta_{rt}(l|s) = \delta\big[l, \operatorname*{argmax}_k \; c_{sk} p(x_{rt}|\lambda_{sk})\big],$$

where $\delta(i,j)$ denotes the Kronecker delta. Accordingly we define the time alignment probability of a word sequence W and of a single word w with word boundary times t_s, t_e in Viterbi approximation [3]:

$$\begin{aligned} s_{rt}(W) &= \operatorname*{argmax}_{s_t} \max_{s_1^{t-1}, s_{t+1}^{T_r}} p(s_1^T, x_{r1}^{T_r}|W), \\ s_{rt}(w,t_s,t_e) &= \operatorname*{argmax}_{s_t} \max_{s_{t_s}^{t-1}, s_{t+1}^{t_e}} p(s_{t_s}^{t_e}, x_{rt_s}^{t_e}|w). \end{aligned}$$

Further we define the *Forward-Backward* (FB) probabilities of the spoken word for word sequence W_r in Viterbi approximation,

$$\begin{aligned} \gamma_{rt}(s;W_r) &= p_\lambda(s_t = s|X_r, W_r) \\ &\overset{\text{Viterbi}}{=} \delta(s, s_{rt}(W_r)) \end{aligned}$$

and the generalized FB probabilities in Viterbi approximation,

$$\begin{aligned} \gamma_{rt}(s) &= \sum_{W \in \mathcal{M}_r} \frac{p^\alpha(X_r, W)}{\sum\limits_{V \in \mathcal{M}_r} p^\alpha(X_r, V)} \gamma_{rt}(s;W) \\ &\overset{\text{Viterbi}}{=} \sum_{t_s,t_e : t_s \le t \le t_e} q_{t_s,t_e}(w|X_r)\, \delta_{s, s_{rt}(w,t_s,t_e)}, \end{aligned} \qquad (1)$$

with

$$q_{t_s,t_e}(w|x_1^T) = \\ = \sum_{\substack{W_s,W_e \\ \in \mathcal{M}}} \frac{p^\alpha(x_1^{t_s-1}|W_s)\, p^\alpha(x_{t_s}^{t_e}|w)\, p^\alpha(x_{t_e+1}^T|W_e)\, p^\alpha(W_s,w,W_e)}{\sum\limits_{V \in \mathcal{M}} p^\alpha(x_1^T|V)\, p^\alpha(V)},$$

where W_s and W_e define starting and ending word sequences enclosing word w. The generalized FB probability $\gamma_{rt}(s)$ denotes the probability for state s at time t, given the total of all alternative word sequences W defined by the sets $\mathcal{M}_r$. Representing these sets by word graphs and using the Viterbi approximation, the sum over all alternative word hypotheses in Eq. (1) could be partially separated from the according time alignments, which leaves us with the sum over time alignments of each word w of the word graph multiplied by the according word probabilities $q_{t_s,t_e}(w|X_r)$ defined above. The time alignments of the words are obtained within the recognition pass. The word probabilities could be calculated efficiently using a forward-backward scheme on word graphs, as described in detail in [10].

Using the above definitions, the density specific discriminative averages are defined by the difference of the averages on the spoken word sequences (spk) and the sets of alternative word sequences represented by the generalized FB probability (gen),

$$\Gamma_{sl}(g(x)) = \Gamma_{sl}^{spk}(g(x)) - \Gamma_{sl}^{gen}(g(x)) \qquad (2)$$

with

$$\begin{aligned} \Gamma_{sl}^{spk}(g(x)) &= \alpha \sum_{r=1}^R f_r \sum_{t=1}^{T_r} \gamma_{rt}(s;W_r) \cdot \eta_{rt}(l|s)\, g(x_{rt}), \\ \Gamma_{sl}^{gen}(g(x)) &= \alpha \sum_{r=1}^R f_r \sum_{t=1}^{T_r} \gamma_{rt}(s) \cdot \eta_{rt}(l|s)\, g(x_{rt}). \end{aligned}$$

Using discriminative averages, derivatives of the discriminative criterion $\mathcal{F}$ with respect to density specific parameters λ_{sl} could be written in the following compact form:

$$\frac{\partial F_D(\lambda)}{\partial \lambda_{sl}} = \Gamma_{sl}\left(\frac{\partial \log c_{sl} p(x|\lambda_{sl})}{\partial \lambda_{sl}}\right). \qquad (3)$$

For further convenience we additionally define the following state specific and global discriminative averages:

$$\begin{aligned} \Gamma_s(g(x)) &= \sum_l \Gamma_{sl}(g(x)) \\ \Gamma(g(x)) &= \sum_s \Gamma_s(g(x)). \end{aligned}$$

Discriminative averages enable to write down reestimation formulae independent of the criterion in use. Accordingly, differences in criteria are introduecd solely by the discriminative averages.

2.1.1. Discriminative Reestimation Formulae

In [6] and [7] we analytically and experimentally showed that parameter optimization of the MMI and the MCE criterion is very similar using the extended Baum-Welch (EB) or gradient descent like methods by introducing a special choice of step sizes for gradient descent. Here we chose the EB optimization method.

Performing the EB algorithm for optimization of the means μ_{sl}, global pooled diagonal variances σ^2 and mixture weights c_{sl} of Gaussian mixture densities, we obtain the following reestimation equations,

$$\begin{aligned} \hat{\mu}_{sl} &= \frac{\Gamma_{sl}(x) + D\, c_{sl} \mu_{sl}}{\Gamma_{sl}(1) + D\, c_{sl}} \\ \hat{\sigma}^2 &= \frac{\Gamma(x^2) + D \sum_s (\sigma^2 + \sum_l c_{sl} \mu_{sl}^2)}{\Gamma(1) + \sum_s \cdot D} \\ &\quad - \sum_s \sum_l \frac{\Gamma_{sl}(1) + D\, c_{sl}}{\Gamma(1) + \sum_s \cdot D} \hat{\mu}_{sl}^2 \\ \hat{c}_{sl} &= \frac{\dfrac{\Gamma_{sl}^{spk}(1)}{\Gamma_s^{spk}(1)} - \dfrac{\Gamma_{sl}^{gen}(1)}{\Gamma_s^{gen}(1)} + C_s}{\sum\limits_{l'} c_{sl'} \left[\dfrac{\Gamma_{sl'}^{spk}(1)}{\Gamma_s^{spk}(1)} - \dfrac{\Gamma_{sl'}^{gen}(1)}{\Gamma_s^{gen}(1)}\right] + C_s} c_{sl}. \end{aligned}$$

For details on the determination of the smoothing parameters D and C_s we refer to [7].

3. DISCRIMINATIVE SPLITTING

As could be seen in Table 4, the improvements obtained by discriminative training methods in comparison to conventional ML training are especially high for single Gaussian density acoustic models, i.e. for low model complexity. On the other hand the relative improvements obtained by discriminative training are reduced for more complex models. The comparatively good performance for low model complexity suggests that discriminative training criteria should be well suited to evaluate the ability of an acoustic model to describe the data. Taking a closer look to the discriminative counts $\Gamma_{sl}(1)$ we come to the following interpretations.

For the case of the MMI criterion, the count $\Gamma_{sl}^{spk}(1)$ for the spoken word sequences just gives the number, how often an observation is aligned to density l and state s given the spoken word sequence. Ideally, the count $\Gamma_{sl}^{gen}(1)$ for the alternative word sequences would be the same, if the posterior probability of the spoken word sequence is always considerably higher than the posterior probabilities of all other word sequences, which suggests suboptimal modeling. Accordingly, if $\Gamma_{sl}^{gen}(1)$ is lower than $\Gamma_{sl}^{spk}(1)$, the spoken word sequence is underrepresented in the set of alternative word sequences. If $\Gamma_{sl}^{gen}(1)$ is higher than $\Gamma_{sl}^{spk}(1)$, then density l in state s even becomes contributions from more alternative word sequences than the spoken ones. Both latter cases suggest sufficiently well modeling. As stated in [4] for the case of the MMI criterion, this suggests that only those densities be splitted, which have the highest values of the discriminative count $\Gamma_{sl}(1)$.

Another heuristic derivation of model evaluation by discriminative averages might be drawn from the derivatives of the unified discriminative criterion by the mixture weights c_{sl} (Eq.(3)),

$$\frac{\partial F_D(\lambda)}{\partial c_{sl}} = \frac{1}{c_{sl}} \cdot \Gamma_{sl}(1).$$

According to gradient descent based parameter optimization, large positive derivatives would indicate large increases in the criterion by increasing the according mixture weight, considering the normalization constraint and provided the criterion is to be maximized. This could also be interpretated as the need of the according density to be better modelled. If, in addition the derivative is multiplied by the according mixture weight itself, i.e. by its relative importance for a given state, we again arrive at the interpretation, that the value of the discriminative count indicates the modeling ability of a density.

After choosing a density for splitting, in conventional splitting the mixture weight is equally distributed upon both new densities and the mean is perturbed by small amounts in opposite directions. In our discriminative splitting approach we instead reestimate the density according ML and MMI to obtain the new pair of densities. In other words, the mean and mixture weight from ML reestimation are assigned to one density, and the mean and mixture weight from MMI reestimation are assigned to the other density, which we believe to be a better estimation than a perturbation around the density to be splitted.

In each splitting step we split 50% of the densities according to their discriminative counts. Finally the resulting increased parameter set is trained until convergence by *Maxmimum Likelihood* (ML) training.

4. EXPERIMENTAL RESULTS

Experiments were performed on the *SieTill* corpus [2] for telephone line recorded German continuous digit strings. The *SieTill* corpus consists of approximately 43k spoken digits in 13k sentences for both training and test.

The recognition system for the *SieTill* corpus is based on whole word HMMs using continuous emission distributions. It is characterized as follows:

- vocabulary of 11 German digits including '*zwo*'
- gender-dependent whole-word HMMs, with every two subsequent states being identical
- for each gender 214 distinct states plus one for silence,
- Gaussian mixture emission distributions,
- global pooled diagonal covariance matrix,
- 12 cepstral features plus first derivatives and the second derivative of the energy.

The baseline recognizer applies ML training using the Viterbi approximation which serves as a starting point for the additional discriminative training. A detailed description of the baseline system could be found in [9].

As shown in Table 4 the best result for conventional splitting with ML training was obtained using 64 densities per mixture, leading to a word error rate of 1.81%.

Table 2: Word error rates on the *SieTill* test corpus obtained for conventional (conv.) mixture density splitting with ML and several discriminative training methods and for discriminative (disc.) splitting with ML and subsequent MMI training. In the column 'dns' the average number of densities per mixture is given.

split criterion	dns	training criterion	error rates [%] del - ins	 WER	 SER
–	1	ML	0.71-0.63	3.78	9.74
		CT	0.76-0.47	2.85	7.27
		MMI	0.81-0.41	2.81	7.13
		FT	0.65-0.64	2.80	7.27
		MCE	0.73-0.41	2.60	6.73
convent.	32	ML	0.46-0.47	1.97	5.31
	64		0.46-0.38	1.81	4.93
	128		0.45-0.39	1.85	4.94
	32	CT	0.52-0.30	1.82	4.97
		MMI	0.42-0.37	1.74	4.80
		FT	0.41-0.37	1.67	4.50
	64	MCE	0.42-0.34	1.69	4.64
discrim.	14	ML	0.42-0.31	1.77	4.77
	33		0.41-0.23	1.61	4.42
		MMI	0.40-0.24	1.61	4.46

The best result for conventional splitting with discriminative training was obtained using only 32 densities per mixture givining a word error rate of 1.67%. The best overall result of 1.61% word error rate on this task we obtained using discriminative splitting and ML training leading to on average 33 densities per mixture. For a solely discriminative splitting approach on the *TI-digitstring* task it was reported in [4] that further ML training gave an increase in error rate. Since our final models are ML trained in the first place, we tried to improve them further by subsequent MMI training, which lead to no additional improvement. Fig. 1 clearly shows, that the results for discriminative splitting are significantly better than those obtained by conventional splitting and both ML and MMI training, especially for equal parameter numbers. Fig. 2 shows a plot of the log-likelihood convergence for ML training with conventional and discriminative splitting. For equal number of parameters, the discriminative splitting approach clearly leads to lower likelihoods than the

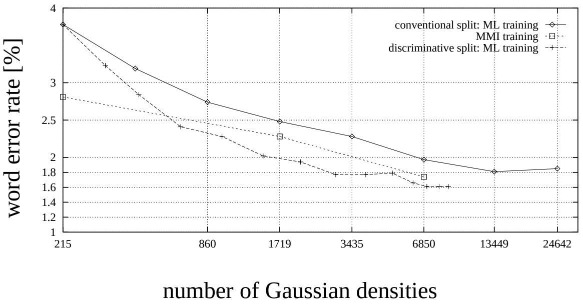

Figure 1: Evolution of word error rates on the *SieTill* test corpus for the proposed combined MMI/ML splitting approach and for conventional splitting with ML and MMI training.

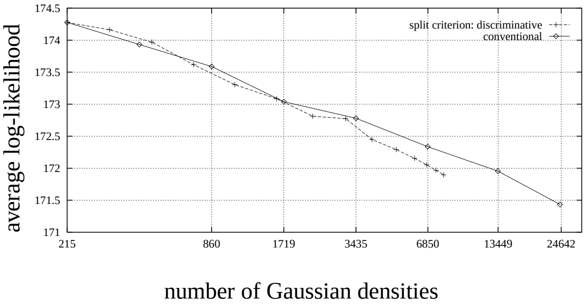

Figure 2: Comparison of the average log-likelihood from ML training against number of Gaussian densities for both splitting approaches considered here (female portion of the *SieTill* corpus).

conventional splitting. Finally, Fig. 3 shows the distribution of numbers of densities for each mixture of the whole word HMM including silence. The values clearly vary very much, ranging from a minimum of 1 density up to 270 densities for the silence mixture. The latter could be motivated by the high overall silence ratio of more than 55% in the *SieTill* corpus.

5. CONCLUSION

In this paper, a combined Maximum Mutual Information (MMI) and Maximum Likelihood (ML) splitting approach was introduced. The MMI criterion was used to evaluate mixture densities for splitting and the ML criterion was used for training the model parameters. Initial word error rates on the *SieTill* corpus for telephone line recorded German continuous digit strings for conventional splitting were 1.81% for ML training and 1.67% for discriminative training. This is to be compared with 1.61% word error rate for discriminative splitting. It should be noted that the result for discriminative splitting with ML training was the same as after further discriminative training.

Currently experiments are performed in order to investigate this splitting approach also for large vocabulary speech recognition.

Acknowledgement. This work was partly supported by Siemens AG, Munich.

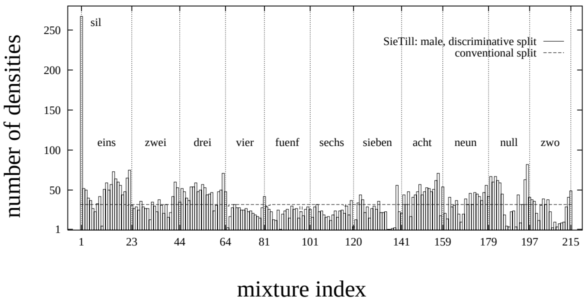

Figure 3: Distribution of the number of densities per mixture obtained by applying the proposed combined MMI/ML splitting approach to the male portion of the *SieTill* corpus.

6. REFERENCES

[1] L. R. Bahl, M. Padmanabhan. “A Discriminant Measure for Model Complexity Adaptation,” Proc. *1998 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, page 453-456, Seattle, WA, May 1998.

[2] T. Eisele, R. Haeb-Umbach, D. Langmann, “A comparative study of linear feature transformation techniques for automatic speech recognition,” in *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, PA, Vol. I, pp. 252-255, October 1996.

[3] H. Ney. “Acoustic Modeling of Phoneme Units for Continuous Speech Recognition,” Proc. *Fifth Europ. Signal Processing Conf.*, Barcelona, pp 65-72, September 1990.

[4] Y. Normandin. “Optimal Splitting of HMM Gaussian Mixture Components with MMIE Training,” Proc. *1995 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, page 449-452, Detroit, MI, May 1995.

[5] Y. Normandin. “Maximum Mutual Information Estimation of Hidden Markov Models,” *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, K. K. Paliwal (eds.), pp. 57-81, Kluwer Academic Publishers, Norwell, MA, 1996.

[6] R. Schlüter, W. Macherey, S. Kanthak, H. Ney, L. Welling. “Comparison of Optimization Methods for Discriminative Training Criteria,” Proc. *1997 Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, Vol. 1, pp. 15-18, September 1997.

[7] R. Schlüter, W. Macherey. “Comparison of Discriminative Training Criteria,” Proc. *1998 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 493-496, Seattle, WA, May 1998.

[8] V. Valtchev, J. J. Odell, P. C. Woodland, S. J. Young. “MMIE training of large vocabulary recognition systems,” Speech Communication, Vol. 22, No. 4, page 303-314, September 1997.

[9] L. Welling, H. Ney, A. Eiden, C. Forbrig. “Connected Digit Recognition using Statistical Template Matching,” Proc. *1995 Europ. Conf. on Speech Communication and Technology*, Madrid, Vol. 2, pp. 1483-1486, September 1995.

[10] F. Wessel, K. Macherey, R. Schlüter. “Using Word Probabilities as Confidence Measures,” Proc. *1998 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 225-228, Seattle, WA, May 1998.