BMBF

Verbmobil

Verbundvorhaben

Discriminative Training for Continuous Speech Recognition

W. Reichl

G. Ruske

Technische Universität München

Vm

Report 110

April 1996

April 1996

W. Reichl
G. Ruske

Forschungsgruppe Sprachverarbeitung
Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
Arcisstraße 21
80290 München

Tel.: (089) 2105 - 8554
e-mail: `reichl@e-technik.tu-muenchen.de`

Gehört zum Antragsabschnitt: TP3 Spracherkennung und Sprecheradaption

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 C/6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

DISCRIMINATIVE TRAINING FOR CONTINUOUS SPEECH RECOGNITION

W. Reichl and G. Ruske

Institute for Human-Machine-Communication,
Munich University of Technology,
Arcisstr. 21, D-80290 München, Germany

ABSTRACT

Discriminative training techniques for Hidden-Markov Models were recently proposed and successfully applied for automatic speech recognition. In this paper a discussion of the Minimum Classification Error and the Maximum Mutual Information objective is presented. An extended reestimation formula is used for the HMM parameter update for both objective functions. The discriminative training methods were utilized in speaker independent phoneme recognition experiments and improved the phoneme recognition rates for both discriminative training techniques.

1. INTRODUCTION

Recently discriminative training techniques for Hidden- Markov Models (HMM) were used successfully for automatic speech recognition. They provide better performance compared to Maximum Likelihood Estimation (MLE), since the training is concentrated on the estimation of class boundaries and not on parameters of assumed model distributions [1,12]. Although MLE and discriminative training are theoretically equivalent (if sufficient classifier parameters and enough training data exist and if Gaussian mixture assumptions are appropriate) discriminative training techniques provide better performance if these requirements are not met [1,12]. A popular alternative to MLE is the Maximum Mutual Information (MMI) between the acoustic observation and the decoded symbols [1,5,9,11,12]. This criterion attempts to minimize the uncertainty about the message, given the observed signal.

Another discriminative objective function is the Minimum Classification Error (MCE), which approximates the misclassification rate of the classifier [3,4,8,13,14]. The optimization of this error function is generally carried out by the Generalized Probabilistic Descent (GPD) algorithm, a gradient descent based optimization, and results in a classifier with minimum error probability [5,8]. In the paper the different objective functions are compared by a uniform formalism. The optimization of the objective functions is carried out by a gradient descent method for the MCE [1,3,4,5,8,9,13,14] or an extended Baum-Welch (BW) algorithm for MMI [9,10,11,12]. In the paper an extended BW algorithm for the MCE criterion is presented, which is faster than a steepest descent based optimization and complies with the constraints for HMM parameters. The discriminative training techniques are used in phoneme recognition experiments with semicontinuous HMMs (SCHMM) and improved the phoneme recognition rates of 5.3 points up to 64.8 % for the MCE optimization.

This work was funded by the German Federal Ministry for Research and Technology (BMFT) in the framework of the Verbmobil Project under Grant 01 IV 102 C6. The responsibility for the contents of this study lies with the authors

2. DISCRIMINANT TRAINING TECHNIQUES FOR SPEECH RECOGNITION

2.1. Minimum Classification Error

Minimum Classification Error (MCE) and Generalized Probabilistic Descent (GPD) have been successfully applied to speech recognition [3,4,8,13,14]. The MCE function is attempting to approximate the misclassification rate of the classifier and its optimization by the GPD algorithm results in a classifier with minimum error [8]. Therefore a generalized distance is used as a discriminance measure $d_c(X)$ between the log score $r_c(X) = \log(p(X|c))$ of the correct model c and the scores $r_n(X)$ of the incorrect models for the acoustic vector sequence $X = \{x_1, \ldots, x_T\}$ with $\eta > 0$ determining the utilized metric:

$$d_c(X) = -r_c(X) + \log\left(\frac{1}{N-1}\sum_{n;n\neq c} e^{r_n(X)\eta}\right)^{\frac{1}{\eta}} \tag{1}$$

$$= \frac{1}{\eta}\log\left(\sum_{n;n\neq c} e^{(r_n(X)-r_c(X))\eta}\right) - \frac{1}{\eta}\log(N-1). \tag{2}$$

The following smoothed 'zero-one' cost function $L(c, X)$ is 'counting' the classification errors and hence approximating the classifier error rate :

$$L(c,X) = l(d_c(X)) = \frac{1}{1+e^{-\gamma d_c(X)}} \text{ with } \gamma > 0. \tag{3}$$

The MCE objective is based on the sigmoid function as a function of the discriminance measure $l(d_c(X))$. Optimization of this continuous objective function with respect to the parameters results in a minimum error classifier [8] and is usually carried out by the GPD algorithm, a general gradient descent optimization, which needs the gradient of the error function with respect to the model scores:

$$\frac{\partial L(c,X)}{\partial r_n(X)} = G_n(X)\, l'(d_c(X)) \tag{4}$$

$$\text{with } \begin{cases} G_n(X) = \frac{(p(X|n))^\eta}{\sum_{n\neq c}(p(X|n))^\eta} & : \ n \neq c \\ G_c(X) = -1 & : \ n = c. \end{cases} \tag{5}$$

Presented at the *Proc. EUROSPEECH-95*, Madrid, Spain, September 1995, pp. 537-540.

The parameters of all HMM models n are updated, using (4), which consists of a model-dependent weighting term $G_n(X)$ and the derivative $l'(d_c(X))$ of the sigmoid function with respect to the misclassification measure $d_c(X)$. It reaches its maximum when the scores of the models are similar and misclassification is likely to occur. For different scores the derivative of the sigmoid function declines rapidly and the GPD training is concentrating on observations, which are likely to be misclassified. In Figure 1 the MCE objective function $L(c,X)$ and the derivative $l'(d_c(X))$ are printed as functions of the score distance $d_c(X)$ for $\gamma = 1$. By the usage of the weights $G_n(X)$ models with higher scores, which are competitors in classification and therefore have to be separated in training, are selected.

In (1) all alternatives of symbol c for X are used. Calculating all possible alternatives of sequences of symbols requires an immense amount of computation power and therefore only the most probable sequence of words or sentences are considered by a 'N-best' search [4]. In the training we use all alternative symbols in the segmentation derived by the correct transcription of the utterance to calculate the misclassification measure (1).

The calculation of the scores $r_n(X)$ can be employed by the Forward-Backward algorithm or a Viterbi decoder. In this paper a Viterbi decoder, based on the most probable state sequence $Q_n = \{Q_{n1}, \ldots, Q_{nT} | Q_{nt} \in \{q_{nm}\}\}$, is used for the calculation of the scores $r_n(X) = \sum_{t=1}^{T} \log p(x_t|Q_{nt})$ of HMM W_n for the acoustic observation $X = \{x_1, \ldots, x_T\}$. Using tied-mixtures HMM, the likelihood $p(x_t|q_{nm})$ of observing vector x_t in state q_{nm} of HMM n is usually a mixture of Gaussians, whereby the Gaussians are shared by all states of all models (SCHMM) or are individually used by states or models in continuous density HMMs. The optimization of the MCE function requires the gradient with respect to the state-specific observation density b_{nm}:

$$\frac{\partial L(c,X)}{\partial b_{nm}} = G_n(X) l'(d_c(X)) \sum_{t:Q_{nt}=q_{nm}} \frac{1}{p(x_t|q_{nm})}. \quad (6)$$

The sum in (6) is considering all times when state q_{nm} is selected by the Viterbi decoder [13]. Similar equations for the Forward-Backward algorithm can be derived.

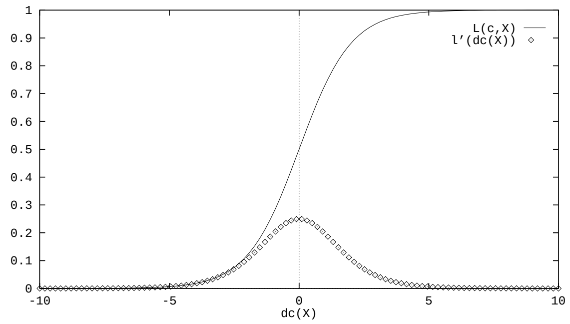

Figure 1: MCE objective function $L(c,X)$ and derivative of the sigmoid function $l'(d_c(X))$ as functions of the misclassification measure $d_c(X)$.

2.2. Maximum Mutual Information

Maximum Mutual Information (MMI) training of HMM classifiers for speech recognition has recently been proposed [1] and successfully applied [1,5,9,11,12]. It attempts to maximize the probability of the correct symbol given the training observation, by maximizing the mutual information $I(c,X)$ between the acoustic observation X and the decoded symbol c, given the a-priori probability of the symbols $p(n)$ [1]. Using the log scores $r_n(X)$ of the models the mutual information can be calculated similar to the MCE function:

$$I(c,X) = \log\left(\frac{p(X|c)}{\sum_n p(n)p(X|n)}\right) \quad (7)$$

$$= r_c(X) - \log\left(\sum_{n=1}^{N} p(n) e^{r_n(X)}\right). \quad (8)$$

The right side of (8) is very similar to the GPD misclassification measure (1) for $\eta = 1$. In the denominator of (8) the probability of the observation $p(X)$ is computed as average of the likelihoods over all possible symbols (including the correct symbol c), weighted by the model priors $p(n)$. The distance measure $d_c(X)$ in the MCE formulation (1) is considering only incorrect models in the summation, according to the utilized metric (η). Optimization of (8) for a sequence of symbols results in the maximization of the correct sequence of symbols c vs all possible sequences of symbols n. Therefore a sentence 'N-best' search [1,5,9] or a general looped model [11,12] is employed to estimate the likelihoods for alternative sequences of symbols. The mutual information $I(c,X)$ is related to the Maximum A-Posteriori decoder, using the a-posteriori probability $p(c|X)$ of the correct symbol: $\log p(c|X) = I(c,X) + \log p(c)$.

Introducing the discriminance measure $d_c(X)$ ($\eta = 1$) between the scores $r_n(X)$ of the models and assuming equal a-priori probabilities for all N classes ($p(n) = \frac{1}{N}$) the mutual information can be calculated by

$$I(c,X) = -\log\left(e^{(d_c(X)+\log(N-1))} + 1\right) + \log(N). \quad (9)$$

The negative mutual information $-I(c,X)$ as a function of the score distance $d_c(X)$ is printed in Figure 2 for $N = 2$. To compare the maximization of the mutual information to the MCE minimization the gradient with respect to the model scores is calculated:

$$\frac{\partial I(c,X)}{\partial r_n(X)} = -G_n(X) l\left(d_c(X) + \log(N-1)\right). \quad (10)$$

The gradient (10) for the MMI training consists of the sigmoid function $l(d_c(X) + \log(N-1))$ with $\gamma = 1$ as function of the shifted distance $d_c(X)$ and the model-dependent weighting terms $G_n(X)$ ($\eta = 1$). Since the MMI objective is maximized $-I(c,X)$ and the sigmoid function for $N = 2$ are printed in Figure 2.

Comparing MMI and MCE objectives we see the MMI functions are not symmetrical. For $d_c(X) < 0$ (i.e. the correct score $r_c(X)$ is higher than the averaged incorrect) both error functions are similar, but for $d_c(X) > 0$, indicating a recognition error, the MMI objective $I(c,X)$ is not bounded. This behavior has some effects in learning, since the gradient of the MCE function consists of the differentiated sigmoid, while the MMI gradient is based on the sigmoid function itself. MCE training is mainly concentrated on the class boundaries, while the sigmoid function in (10) is putting emphasis on extreme false classifications ($d_c(X) \gg 0$). This can cause problems in the training, which is highly influenced by outlayers (e.g. due to incorrect labeling).

The maximization of the mutual information of tied-mixtures HMM using the Viterbi algorithm requires the gradient with respect to the state-specific observation density b_{nm}:

$$\frac{\partial I(c,X)}{\partial b_{nm}} = (\delta_{nc} - p(n|X)) \sum_{t:Q_{nt}=q_{nm}} \frac{1}{p(x_t|q_{nm})}, \qquad (11)$$

whereby δ_{nc} is the Kronecker delta. In (11) a probabilistic interpretation of (10) is given, which consists of the difference between the 'desired' and the actual a-posteriori probability of symbol n. The required terms for the HMM training with the Forward-Backward algorithm are presented in [11,12].

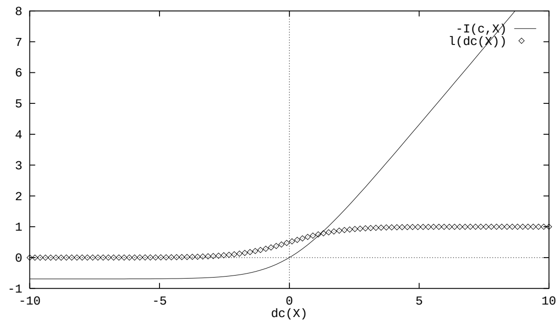

Figure 2: Negative mutual information $-I(c,X)$ and sigmoid function $l(d_c(X))$ as function of the distance $d_c(X)$ for $N = 2$.

2.3. Optimization Techniques

Usually HMM learning is based on the Maximum Likelihood principle, optimizing the likelihood of the observation by a very efficient parameter reestimation technique, the Baum-Welch (BW) algorithm [3,7]. The optimization of HMM parameters according to discriminative criteria may be carried out with standard optimization techniques, such as steepest descent or conjugate gradients [1,3,5,8,9, 13,14]. Since some HMM parameters λ_i are constrained (e.g. $\sum_i \lambda_i = 1$), an additional parameter transformation is required to meet the Lagrange conditions for the constraints [3,7,9,14].

In [6] an improved BW algorithm for the training of rational functions R(X) (e.g. $I(c,X)$) was presented, which was extended in [11,12] to continuous observation densities. In speech recognition experiments this extended BW algorithm showed improved convergence compared to gradient descent training [9,11,12]. Recently the theoretical conditions for objective functions, optimized by (12), were relaxed to general analytic functions [10] (e.g. $L(c,X)$). The reestimation formula for parameter λ_i is very similar to the BW growth-transformation [2]:

$$\hat{\lambda}_i = T^D(\lambda_i) = \frac{\lambda_i \left(\frac{\partial R(X)}{\partial \lambda_i} + D\right)}{\sum_i \lambda_i \left(\frac{\partial R(X)}{\partial \lambda_i} + D\right)}. \qquad (12)$$

Update formulas for special parameters, such as Gaussian means, are printed in [11,12]. The growth-transformation $T^D(\lambda_i)$ can be reduced to the original BW transformation $T^0(\lambda_i)$ ($D = 0$): $T^D(\lambda_i) = \beta(D)T^0(\lambda_i) + (1-\beta(D))\lambda_i$ with

$$\beta(D) = \frac{\sum_i \lambda_i \frac{\partial R(X)}{\partial \lambda_i}}{\sum_i \lambda_i \frac{\partial R(X)}{\partial \lambda_i} + D}. \qquad (13)$$

The convergence of the growth-transformation $T^D(\lambda_i)$ for any analytic function is ensured with $0 \le \beta < \beta(D)$ for discrete observations [10]. Although for continuous observation densities only $D \to \infty$ theoretically ensures convergence [11], MMI training with

$$D(X) = \max_{\lambda_i} \left\{ -\frac{\partial R(X)}{\partial \lambda_i}, 0 \right\} + \epsilon \qquad (14)$$

showed fast, but not strict monoton learning [9,11,12]. This can be further improved by parameter smoothing, which we used for the MMI and MCE training of SCHMMs:

$$T^D(\lambda_i) = \alpha\beta(D)T^0(\lambda_i) + (1 - \alpha\beta(D))\lambda_i. \qquad (15)$$

The smoothing parameter $\alpha : \{\alpha | 0 < \alpha \le 1\}$ controls the degree of parameter change, similar to the general step-size in gradient descent, while the parameter dependent term $\beta(D)$ performs a data dependent control of the step-size. According to [7,10], every smoothing parameter $0 < \alpha \le 1$ in the original BW algorithm ($D = 0 : T^0(\lambda_i)$) results in an increasing likelihood, while for different objective functions α is to be determined. Using (15) the new parameters λ_i are restricted to meet the Lagrange conditions and the change of parameters has a positive projection along the gradient of $R(X)$, which is a condition for the optimization of $R(X)$.

3. EXPERIMENTS

MCE and the MMI training techniques were applied in speaker-independent phoneme recognition experiments to compare the convergence and performance of the different discriminative approaches. About 7700 sentences from 67 speakers from a German database with continuous speech (Phondat "Diphon") were utilized for the training of 41 phoneme models with 3 to 6 states for each phoneme and about 3300 sentences from 33 other speakers for the test. The speech data were sampled at 16kHz, and a 256-point FFT with Hamming window was calculated every 10ms to compute the normalized loudness spectrum in 20 critical (Bark-scaled) bands. The delta-loudness spectrum and the total loudness together with the zero-crossing rate were added as separate features to the SCHMM 'soft' vector quantization. The individual features were processed in separate codebooks with 256, 128 and 32 Gaussian pdfs with diagonal covariance matrices, derived from the LBG-clustering.

The SCHMMs are first optimized according to the MLE principle by a Viterbi training algorithm. Phoneme recognition rates were evaluated within an automatically determined phoneme segmentation and resulted in a phoneme recognition performance for the test data of 59.5 % for the baseline system.

In the following optimization of the discriminative functions the extended BW algorithm (12) was applied to reestimate the mixtures of the phoneme models. All phoneme alternatives within the automatically derived phoneme segmentation were used in the calculation of the discriminance measure. According to (6), the mixtures of the correct and of all competing models were updated to minimize the objective function. Best results were obtained in MCE optimization using the update equation (15) with $\alpha = 1.0$, which results in a stable learning and a phoneme recognition rate of 64.8 % on the test set. For $\alpha = 0.3$ objective function and error rate were decreasing monotonously, but recognition performance within the 5 iterations was slightly worse (62.9 %).

MMI training, using the extended BW algorithm, was stable for a smoothing parameter $\alpha = 0.3$, in which the objective function and the recognition rate were increasing monotonously. The MMI objective is less robust than the MCE function and therefore requires smaller 'step-sizes'. 62.0 % phoneme recognition rate were achieved by the MMI optimization, which is an improvement of 2.5 points compared to the ML baseline system. In these experiments MCE training was more stable than MMI optimization and resulted in higher phoneme recognition results.

Furthermore an alternative calculation of the discriminance measure based on the best phoneme sequence hypothesis for the utterance was applied. Now the correct description versus the best hypothesis of the utterance, which was derived by a looped phoneme model without a lexicon or language model, was used to compute the objective function. Since both descriptions differ only in some parts of the sentence, just these different segments were actually used in discriminative training. Therefore only small parts of the training data were used in the discriminative parameter reestimation process. Only MCE optimization was examined for this technique, which resulted in minor improvements of 1.7 points to 61.2 % phoneme recognition rate within the same number of iterations. Using 'N-best' alternative hypothesis would improve this training scheme by generating more competing hypotheses for the discriminative training.

4. CONCLUSIONS

In this paper a discussion of the MCE and the MMI objective was presented. For the HMM parameter update an extended BW reestimation formula was suggested, which can be used for both discriminative methods. It was applied in speaker independent phoneme recognition experiments and improved the recognition rate about 5.3 points from 59.5 % to 64.8 % for the MCE function. In our experiments MCE training was more stable and resulted in better performance than MMI learning under identical conditions. Since the extended BW algorithm was only applied to HMM mixture coefficients it will be used for MCE optimization of pdf parameters in future, which will further improve the performance of the decoder.

5. REFERENCES

[1] L.Bahl, P.Brown, P.deSouza, R.Mercer, *Maximum Mutual Information Estimation of Hidden Markov Parameters for Speech Recognition*, ICASSP 1986, Tokyo, pp. 49-52, April 1986.

[2] L.E.Baum, J.A.Eaton, *An Inequality with Applications to Statistical Prediction for Functions of Markov Processes and to Model Ecology*, Bull. Amer. Math. Soc., vol. 73, pp. 360-363, 1967.

[3] W.Chou, B.H.Juang, C.H.Lee, *Segmental GPD Training Of HMM Based Speech Recognizer*, ICASSP 1992, San Francisco, pp. 473-476, March 1992.

[4] W.Chou, C.H.Lee, B.H.Juang, *Minimum Error Rate Training Based On N-Best String Models*, ICASSP 1993, Minneapolis, pp. 652-655, April 1993.

[5] Y.L.Chow, *Maximum Mutual Information Estimation Of HMM Parameters For Continuous Speech Recognition Using The N-Best Algorithm*, ICASSP 1990, Albuquerque, pp. 701-704, April 1990.

[6] P.Gopalakrishnan, D.Kanevsky, A.Nadas, D.Nahamoo, *An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems*, Trans. on Information Theory, vol. 37, no. 1, pp. 107-113, July 1991.

[7] S.E.Levinson, L.R.Rabiner, M.M.Sondhi, *An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition*, Bell System Technical Journal, vol. 62, no. 4, April 1983.

[8] B.H.Juang, S.Katagiri, *Discriminative Learning for Minimum Error Classification*, Trans. on Signal Processing, vol. 40, no. 12, pp. 3043-3054, Dec. 1992.

[9] S.Kapadia, V.Valtchev, S.Young, *MMI Training For Continuous Phoneme Recognition On The Timit Database*, ICASSP 1993, Minneapolis, pp. 491-494, April 1993.

[10] D.Kanevsky, *A Generalization Of The Baum Algorithm To Functions On Non-Linear Manifolds*, ICASSP 1995, Detroit, pp. 473-476, May 1995.

[11] Y.Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem*, Ph.D. Thesis, McGill University, Montreal, June 1991.

[12] Y.Normandin, R.Cardin, R.DeMori, *High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation*, Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994.

[13] W. Reichl, P. Caspary, G. Ruske, *A New Model-Discriminant Training Algorithm For Hybrid NN-HMM Systems*, ICASSP 1994, Adelaide, pp. 677-680, April 1994.

[14] W.Reichl, G.Ruske, *A Hybrid RBF-HMM System for Continuous Speech Recognition*, ICASSP 1995, Detroit, pp. 3335-3338, May 1995.