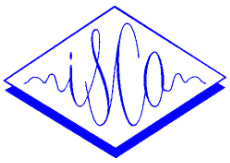
A NOVEL LOSS FUNCTION FOR THE OVERALL RISK CRITERION BASED DISCRIMINATIVE TRAINING OF HMM MODELS

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Janez Kaiser, Bogomir Horvat, Zdravko Kačič

University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, SI-2000 Maribor, Slovenia
e-mail: {janez.kaiser,bogo.horvat,kacic}@uni-mb.si



ABSTRACT

In this paper, [1] we propose a novel loss function for the overall risk criterion estimation of hidden Markov models. For continuous speech recognition, the overall risk criterion estimation with the proposed loss function aims to directly maximise word recognition accuracy on the training database. We propose reestimation equations for the HMM parameters, which are derived using the Extended Baum-Welch algorithm. Using HMM, trained with the proposed method, a decrease of word recognition error rate of up to 17.3% has been achieved for the phoneme recognition task on the TIMIT database.

1. INTRODUCTION

Most of the current automatic speech recognisers use the maximum a posteriori (MAP) decoder, which selects the sentence $\hat{w}$, given the acoustic observation $\mathbf{o}$, according to the following Bayes decision rule [2]:

$$\hat{w} = \underset{w}{\operatorname{argmax}}\, P(w|\mathbf{o}) = \underset{w}{\operatorname{argmax}}\, P(\mathbf{o}|w)P(w). \qquad (1)$$

The acoustic model provides the conditional probability $P(\mathbf{o}|w)$ and the language model provides the a priori probability $P(w)$. When recognising natural speech, the true distributions of $P(\mathbf{o}|w)$ and $P(w)$ are not known. Therefore, we must choose some parametric representation of these distributions. Currently, the most popular parametric representation of the acoustic model probability is based on Hidden Markov Models (HMM).

During the training of a recogniser, based on the HMM models, the aim is to estimate such a set of HMM parameters Θ, which will result in the lowest possible recognition error rate. The most popular method for the estimation of the HMM parameters is Maximum Likelihood Estimation (MLE). Given R acoustic observations in the training database, the objective function used in MLE is

$$f_{MLE}(\Theta) = \prod_{r=1}^{R} P(\mathbf{o}_r|\Theta_r). \qquad (2)$$

[1]This work was funded by the Ministry of Science and Technology, Slovenia, under the contract number 3411-98-22-0854.

Here, $P(\mathbf{o}_r|\Theta_r)$ is the HMM parametric representation of the $P(\mathbf{o}_r|w_r)$.

Nadas [9] has shown that given some restrictions, MLE will produce the best possible decoder. However, these restrictions are almost never met in practical speech recognition applications. Consequently, alternative methods for the HMM parameter estimation have been proposed and successfully implemented in recent years. Two of the most frequently used are Maximum Mutual Information Estimation (MMIE) [1, 10, 11] and Minimum Classification Error Estimation (MCEE) [6, 5].

More recently, Na et al. [8] proposed a new training method, to which we will refer as the Overall Risk Criterion Estimation (ORCE) throughout this paper. It was derived directly from the minimum-error rate classification of the Bayes decision theory. The objective function in ORCE as proposed by Na et al. is

$$f_{ORCE}(\Theta) = \sum_{r=1}^{R} (1 - P(w_r|\mathbf{o}_r))p(\mathbf{o}_r). \qquad (3)$$

In the experiment, described in [8], usefulness of ORCE was demonstrated on the isolated digits recognition task.

In this paper, we extend the work done by Na et al. to the case of continuous speech recognition. Instead of using the zero-one loss function as in [8] for the derivation of the ORCE objective function, we propose a new loss function, which depends on the number of errors (substitutions, insertions and deletions).

2. OVERALL RISK CRITERION

Let $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ be a finite set of M possible sentences and $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ be a finite set of M possible actions. Here action α_i is to select sentence w_i as the recognition result, given an acoustic observation $\mathbf{o}$. We assume that the only possible recognition result is one of the M sentences (we do not allow reject result). Let $\lambda(\alpha_i|w_j)$ be the loss incurred for taking action α_i when the true sentence is w_j. The expected loss associated with the action α_i, given an acoustic observation $\mathbf{o}$, is

$$\mathcal{R}(\alpha_i|\mathbf{o}) = \sum_{m=1}^{M} \lambda(\alpha_i|w_m)P(w_m|\mathbf{o}). \qquad (4)$$

Let us assume that the decision function $\alpha(\mathbf{o})$ is given (in our case, it is the Bayes decision rule, defined in equation (1)). Then, the overall risk is

$$\mathcal{R} = \int \mathcal{R}(\alpha(\mathbf{o})|\mathbf{o})p(\mathbf{o})d\mathbf{o}, \tag{5}$$

where the integral extends over the whole acoustic observation space. For the limited number of R acoustic observations in the training database, we can rewrite the expression (5) as:

$$\begin{aligned}\mathcal{R} &= \sum_{r=1}^{R} \mathcal{R}(\alpha(\mathbf{o}_r)|\mathbf{o}_r)p(\mathbf{o}_r) \\ &= \sum_{r=1}^{R}\sum_{m=1}^{M} \lambda(\alpha(\mathbf{o}_r)|w_m)P(w_m|\mathbf{o}_r)p(\mathbf{o}_r). \end{aligned} \tag{6}$$

We assume the $p(\mathbf{o}_r)$ to be uniform, since it does not depend on the parameter set Θ [8]. By applying the Bayes rule, and replacing $P(\mathbf{o}_r|w_r)$ with its HMM parametrisation $P(\mathbf{o}_r|\Theta_r)$, we arrive at the final expression for overall risk criterion:

$$f_{ORCE}(\Theta) = \sum_{r=1}^{R} \frac{\sum_{m=1}^{M} \lambda(\alpha(\mathbf{o}_r)|w_m)P(\mathbf{o}_r|\Theta_m)P(w_m)}{\sum_{m=1}^{M} P(\mathbf{o}_r|\Theta_m)P(w_m)}. \tag{7}$$

What remains open in the equation (7) is the selection of the loss function $\lambda(\cdot)$. Most often, a so-called symmetrical or zero-one loss function is chosen, which assigns no loss to a correct decision and unit loss to any error [2]:

$$\lambda(\alpha_i|w_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i,j = 1,\ldots,M. \tag{8}$$

Using this loss function, Na et al. [8] derived the objective function, defined in equation (3).

All errors are equally costly when the zero-one loss function is used. This is a reasonable choice for isolated word recognition. In continuous speech recognition, the usual measure of recognition performance is the word recognition accuracy, which depends on the number of substituted S, inserted I and deleted D words in each sentence:

$$\%Accuracy = \frac{N - S - I - D}{N}. \tag{9}$$

Here, N is the total number of words in the recognised sentence.

In the case of continuous speech, the use of zero-one loss function recognition in equation (7) would result in equal loss of one being assigned to all incorrect sequences, regardless of the number of substituted, inserted and deleted words. For this reason, we propose a different loss function, which penalises the incorrect sequences according to the number of errors in them and is in our opinion more closely correlated to the recognition accuracy as the zero-one loss function. The proposed loss function is given by:

$$\lambda(\alpha_i|w_j) = \begin{cases} 0 & i = j \\ e(i,j) & i \neq j \end{cases} \quad i,j = 1,\ldots,M. \tag{10}$$

Here $e(i,j)$ is the number of errors (substitutions, insertions and deletions), computed when w_i is compared to w_j. $e(i,j)$ can be computed using the dynamic programming algorithm which tries to match w_i and w_j. Contrary to the zero-one loss function, which makes no distinction between incorrect sentences, proposed loss function penalises sentences according to their incorrectness in the word recognition accuracy sense. If the objective function (7) with the proposed loss function is minimised on a given training set, the word recognition accuracy on this set (and hopefully also on the test set) should increase.

3. TRAINING ALGORITHM

$f_{ORCE}(\Theta)$, defined in equation (7), is a rational function with the form $S_1(\Theta)/S_2(\Theta)$, where $S_1(\Theta)$ and $S_2(\Theta)$ are polynomials with real coefficients and variables θ_j defined in the domain $D : \theta_j \geq 0, \sum_{j=1}^{q} \theta_j = 1$. Also, $S_2(\Theta)$ has only positive values in the domain D. Because of these properties, $f_{ORCE}(\Theta)$ can be optimised using extended Baum-Welch algorithm (EBW) [3]. New estimates of the parameters θ_j are given by the following transformation:

$$\hat{\theta}_j = \frac{\theta_j(\frac{\partial f_{ORCE}(\Theta)}{\partial \theta_j} + C)}{\sum_{k=1}^{q} \theta_k \frac{\partial f_{ORCE}(\Theta)}{\partial \theta_k} + C}. \tag{11}$$

The values of the constant C can be determined using the formula proposed in [3]:

$$C = \max_{\theta}\left\{-\frac{\partial R((\Theta)}{\theta}, 0\right\} + \epsilon. \tag{12}$$

The partial derivatives of $f_{ORCE}(\Theta)$, needed for the EBW formula, are given by

$$\frac{\partial f_{ORCE}(\Theta)}{\partial \theta_j} = \sum_{r=1}^{R}\sum_{m=1}^{M} K_{r,m}\frac{\partial P(\mathbf{o}_r|\Theta_m)}{\partial \theta_j}, \tag{13}$$

where

$$\begin{aligned} K_{r,m} = & \frac{\left(\sum_{n=1}^{M} P(\mathbf{o}_r|\Theta_n)P(w_n)\right)\lambda(\alpha(\mathbf{o}_r)|w_m)}{\left[\sum_{n=1}^{M} P(\mathbf{o}_r|\Theta_n)P(w_n)\right]^2} \\ & - \frac{\left(\sum_{n=1}^{M} \lambda(\alpha(\mathbf{o}_r)|w_n)P(\mathbf{o}_r|\Theta_n)P(w_n)\right)P(w_m)}{\left[\sum_{n=1}^{M} P(\mathbf{o}_r|\Theta_n)P(w_n)\right]^2}. \end{aligned} \tag{14}$$

Partial derivatives $\frac{\partial}{\partial \theta_j} P(\mathbf{o}_r|\Theta_m)$ with respect to different HMM parameters can be computed using usual formulas (see for instance [4]).

EBW formula is valid only for HMM parameters from a finite set. For the estimation of continuous density parameters (means and variances), we used an extension of the EBW formula, equivalent to the one proposed in [10].

4. EXPERIMENTS

We have tested the performance of the proposed ORCE criterion in comparison with the MLE criterion on the phoneme recognition task using the TIMIT database. Training material for both criteria consisted of 3696 sentences from the SI an SX part of the TRAIN corpus. For the test material, the whole TEST corpus was used, comprising 1344 sentences.

The choice of phonemes for the experiments was similar to work in [7]. The original set of 61 phonemes was mapped to set of 48, which was used during training. During scoring, within-group confusions in seven groups of phonemes were not counted. Thus, we effectively had 39 phones in separate categories.

The speech signal was first pre-emphasised using the filter with the transfer function $s'_n = s_n - 0.95 s_{n-1}$. Signal was then windowed every $10ms$ using Hamming window with length of $25ms$. 39 dimensional feature vectors were used, consisting of 12 mel frequency cepstral coefficients and window energy together with their first and second derivatives.

Phonemes were modelled with context-independent, left-right HMM with 3 emitting states and no skip transitions. Continuous output distributions were used, consisting of mixtures of 1, 2, 4 and 8 Gaussian densities, respectively.

HMM parameters were first optimised with the MLE, using the Baum-Welch algorithm. Parameters from the iteration that produced the highest recognition accuracy were used as the starting parameters for the ORCE training. For the M competing phoneme sequences, the 10 highest-probability sequences were used for each acoustic observation. These were produced using n-best recognition using phoneme lattices.

Phoneme sequences were recognised using a bigram language model, which was trained on the transcriptions of the HMM training corpus. The language model probabilities were raised to the power of 2.0.

Figure 1 shows how the value of the ORCE train criterion evolved over the first 50 iterations for the 1-mixture system. It can be seen that a reasonably stable convergence has been achieved as the value steadily decreased without larger oscillations.

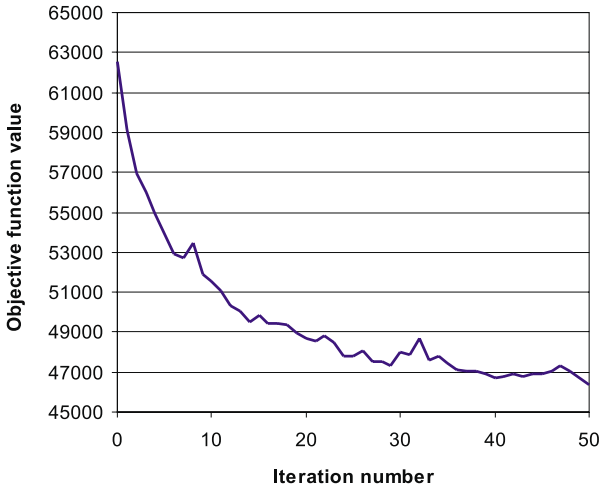


Figure 1: Convergence of the ORCE training criterion in the 1-mixture system over the first 50 iterations.

Figure 2 shows the evolution of word recognition results for the ORCE training criterion over first 50 iterations of training algorithm for the same system. A strong correlation of the recognition results and the ORCE criterion (see Figure 1) can be observed.

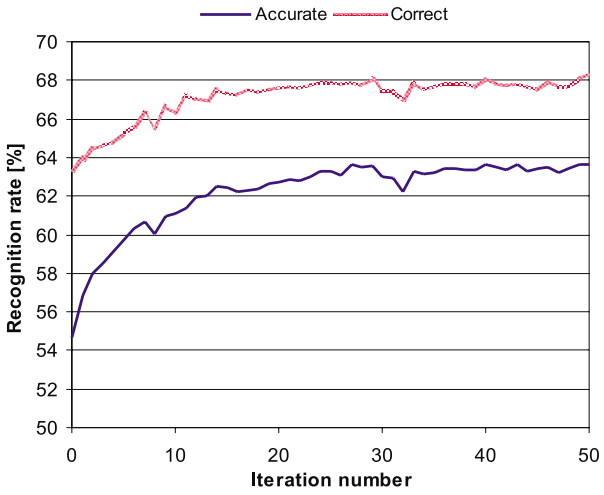


Figure 2: Word recognition results for the ORCE trained 1-mixture system over the first 50 iterations.

Table 1 summarises the best word recognition accuracy results achieved using MLE and ORCE criteria. ORCE has provided a significant and consistent improvement in recognition accuracy and correctness. The recognition error was reduced by 17.3% in the 1-mixture system and by 10.8% in the 8-mixture system. The word recognition accuracy achieved using the 1-mixture ORCE trained system is higher than accuracy, achieved using the 4-mixture MLE trained system.

Type	MLE		ORCE	
	Corr	Acc	Corr	Acc
1-mix	63.33%	54.64%	67.45%	62.50%
2-mix	66.64%	58.93%	69.89%	64.96%
4-mix	68.85%	61.94%	71.25%	66.68%
8-mix	70.69%	64.39%	72.71%	68.25%

Table 1: Comparison of word recognition results for MLE and ORCE training criteria.

5. CONCLUSION

In this paper, we have proposed a new loss function for the overall risk criterion (ORCE) discriminative training of HMM models. The loss function is designed for the case of continuous speech recognition, where three types of errors occur: substitutions, insertions and deletions. Using the ORCE criterion, we have developed reestimation formulas, based on the EBW algorithm. When applied to the phoneme recognition task on the TIMIT database, the use of the proposed reestimation formulas results in a stable convergence of the ORCE criterion and a significant reduction of the word recognition error rate. The highest improvements were achieved for the 1-mixture system, where the word recognition performance has increased from 63.33% correct / 54.64% accurate with the MLE trained HMM to 67.45% correct / 62.50% accurate with the ORCE trained HMM.

6. REFERENCES

1. L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In *Proc. ICASSP '86*, pages 49–52, 1986.
2. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.
3. P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo. A Generalization of the Baum Algorithm to Rational Objective Functions. In *Proc. ICASSP '89*, pages 631–634, 1989.
4. X.D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech recognition.* Edinburgh University Press, 1990.
5. B.-H. Juang, W. Chou, and C.-H. Lee. Minimum Classification Error Rate Methods for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):257–265, December 1997.
6. B.-H. Juang and S. Katagiri. Discriminative Learning for Minimum Error Classification. *IEEE Transactions on Signal Processing*, 40(12):3043–3054, December 1992.
7. K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, November 1989.
8. K. Na, B. Jeon, D. Chang, S. Chae, and S. Ann. Discriminative Training of Hidden Markov Models using Overall Risk Criterion and Reduced Gradient Method. In *Proc. Eurospeech'95*, pages 97–100, 1995.
9. Arthur Nadas. A Decision-Theoretic Formulation of a Training Problem in Speech Recognition and a Comparison of Training by Unconditional Versus Conditional Maximum Likelihood. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-31(4):814–817, August 1983.
10. Y. Normandin. *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem.* PhD thesis, McGill University, Montreal, 1991.
11. V. Valtchev, J.J. Odell, P.C. Woodland, and S.J. Young. Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition. In *Proc. ICASSP '96*, 1996.