Global Discriminative Training of a Hybrid Speech Recognizer

Carla Lopes [1,2,3], *Fernando Perdigão* [1,2]

[1] Instituto de Telecomunicações, [2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal, [3] Instituto Politécnico de Leiria-ESTG, Portugal

{calopes, fp}@co.it.pt

Abstract

Hybrid speech recognizers usually involve a frame-based classification followed by a segment alignment system, trained separately. The simplicity of such systems is counterbalanced by the lack of a global optimisation scheme for the whole system. In this paper we propose a discriminative training method for MLP/HMM hybrids based on the optimization of a global cost function at the phone recognition level. The MLP weights, usually updated according to the target values, are now updated according to the misclassifications present in the output of the system. Results are presented for the TIMIT phone recognition task and show that this method compares favourable with recent published results in this task. The global discriminative training method was also applied to a Portuguese speech database leading to promising results.

Index Terms: discriminative training, hybrid speech recognizers, phone recognition.

1. Introduction

Hybrid speech recognizers have been used with considerable success in several applications, [1-7]. The hybrid framework, in which discriminative classifiers (like Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Conditional Random Fields (CRFs)) are combined with generative models (Hidden Markov Models (HMMs)), allowed for significant gain in performance with respect to standard HMM in several situations. Discriminative training approaches, which are also applied to HMMs systems, aim to minimize the training error. Different training criteria have been successfully tested: Maximum Mutual Information (MMI),[8],[9] Minimum Classification Error (MCE) [10], Minimum Phone/Word Error (MPE/MWE) [11], and methods based on the Principle of Large Margin (PLM), [12]. However, training such a hybrid system is not straightforward, and that justifies why usually classification and alignment undertake separate training steps.

Most hybrid systems are prone to inferior performance due to the lack of a global optimisation scheme for the whole system. One of the persisting challenges is, therefore, to design an integrated discriminative training method to train hybrid systems as a whole. Bengio *et al*, [2] have already focused on this goal, proposing a hybrid system where eight ANN outputs (classifying plosives) were used as inputs for an HMM system, whose states were modulated by Gaussian Mixtures Models (GMMs). Droppo and Acero, [8], proposed a general discriminative training method but applied to both the front-end feature extractor and back-end acoustic model of an automatic speech recognition system. Wu and Huo in [6], propose a MCE training approach for the joint design of a feature compensation module (SVM) and HMM parameters of a speech recognizer. In [13], Riis proposes a hybrid ANN/HMM, called Hidden Neural Networks, in which all parameters are estimated simultaneously according to the discriminative conditional maximum likelihood (CML) criterion. The approach proposed in this paper is somehow related to the state-corrective CML (SCCML) method described by Johansen, [7] used in the computation of a free grammar gradient, now extended to the train of a hybrid recognizer.

In this paper we propose a discriminative training method applied to a hybrid ANN/HMM phone recognizer. The ANN consists of a Multi Layer Perceptron (MLP) network, whose frame-based outputs represents a posteriori probabilities of phone occurrences and are used as state occupancy probabilities in HMMs. A global backpropagation learning scheme is defined considering a strict integration between the HMM and the ANN. The error minimization is based on the gradient descent algorithm and the result is a maximization of the phone accuracy rate and not likelihood maximization, as usual. Phones models are trained to maximize their accuracy rate whilst also maximizing the distance between the correct phone and its rivals. The main goal is to improve phone accuracy in the aligned output string, instead of in the Multi Layer Perceptron output, as usually done. The method uses the difference between the reference and the best acoustic likelihood of the observation sequences to update the MLP weights.

2. Global Discriminative Training Method

A global discriminative training method (GDTM) for training the parameters of a hybrid MLP/HMM as a whole is proposed. MLP is a natural structure for discriminative training; however the network weights are usually updated according to the target values presented in the output layer rather than according to the best sequence of HMM states. To overcome this problem, we propose a training method based on a cost function that minimizes the classification error of the global hybrid system, operating at the recognition level. The free parameters of the system are updated according to the misclassifications between the labeled sequence and the reference. Figure 1 illustrates the proposed method.

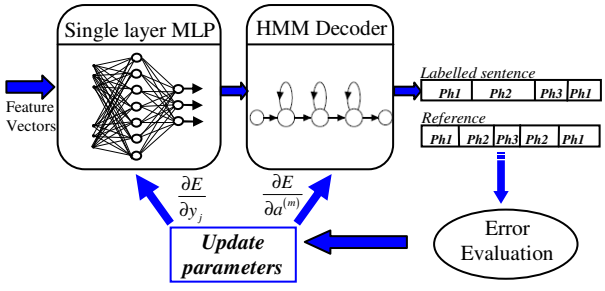

Figure 1: *Schematic diagram of the proposed global discriminative training method.*

The goal is to compute the gradient of the cost function with respect to the MLP outputs and to backpropagate this gradient through the entire structure, all the way back to the first MLP layer. The output alignment (Viterbi trellis) plays a major role in the training process, since the gradient of the cost function, with respect to the MLP outputs is computed based on this alignment. In order to have the best alignment sequence to compare with the reference alignment, a Viterbi decoder was fully incorporated into our training scheme.

2.1. Cost Function

Formulation and estimation of a correctly specified cost function would be central to the discriminative training procedure. The cost function should focus on multiple decoding alternatives, for instance using a N-best list, and consider all kinds of errors: substitutions, insertions and deletions. In a first approach, we used only the contribution of the best hypothesis provided by the Viterbi decoder. In this case the *Levenshtein* distance aligns two label sequences. One is the correct sequence, W_{lab}, and the other is the best decoding hypothesis given by the recognizer, W_{rec}. Using the Viterbi algorithm, we define an error function as:

$$d\left(W_{rec}, W_{lab}\right) = g(W_{rec}) - g(W_{lab}) \tag{1}$$

where $g(W_{lab})$ and $g(W_{rec})$ represent the reference and the best acoustic likelihood of the observation sequence. This difference is always greater than zero, and is only zero if the two transcriptions are exactly the same (labels and time alignments coinciding).

If N_{BD} is the total number of training utterances, the global cost is then given by:

$$E = \sum_{n=1}^{N_{BD}} d(W_{rec}^{(n)}, W_{lab}^{(n)}) = \sum_{n=1}^{N_{BD}} e^{(n)} . \tag{2}$$

If $W = W_1^{N_W} = \{w_1, w_2, \cdots, w_{N_W}\}$ is the sequence of phones in an utterance, the total log-likelihood (assuming a bigram model) is given by:

$$g(W) = \sum_{k=1}^{N_W} \left[\log\left(\mathrm{P}\left(X_{t_{k-1}}^{t_k} \mid w_k\right)\right) + \log\left(\mathrm{P}(w_k \mid w_{k-1})\right) \right] \tag{3}$$

where

$$\log\left(\mathrm{P}\left(X_{t_{k-1}}^{t_k} \mid w_k\right)\right) =$$

$$= \sum_{t=t_{k-1}}^{t_k} \left(\log\left(a_{s_{t-1}, s_t}\right) + \log\left(b_{s_t}(\mathbf{x}_t)\right) \right) + \log\left(a_{s_{t_k}, s_{t_k}+1}\right)$$

is the cost of traversing the HMM of phone w_k with observations from $t=t_{k-1}$ to $t=t_k$. The function $b_s(\mathbf{x})$ is the likelihood of observing $\mathbf{x}$ in the HMM state s. The last term in the previous equation corresponds to the exit probability of the w_k HMM.

In the hybrid system the MLP output predictions are interpreted as the a posteriori phone probabilities of jth phone/state, $\mathrm{P}\left(s_j \mid \mathbf{x}\right)$, given the feature observation vector x. The likelihood ratio, $\mathrm{P}\left(\mathbf{x} \mid s_j\right) / \mathrm{P}(\mathbf{x})$, used in the HMM framework, is replaced by the posterior probabilities, using Bayes's rule,

$$\frac{\mathrm{P}\left(\mathbf{x} \mid s_j\right)}{\mathrm{P}\left(\mathbf{x}\right)} = \frac{\mathrm{P}\left(s_j \mid \mathbf{x}\right)}{\mathrm{P}\left(s_j\right)} . \tag{4}$$

The a priori phone probabilities $\mathrm{P}(s_j)$ are estimated off-line from the training data.

2.2. Gradient with respect to the outputs of the MLP

We used the gradient descendent method to update the network weights. In this case the error gradient for an MLP output y_j is:

$$\frac{\partial E}{\partial y_j} = \sum_{n=1}^{N_{BD}} \frac{\partial e^{(n)}}{\partial y_j}$$

$$\frac{\partial e^{(n)}}{\partial y_j} = \left(\sum_{t=1}^{T_n} \delta\left(j, s_t^{(rec)}\right) \frac{1}{y_j} \right) - \left(\sum_{t=1}^{T_n} \delta\left(j, s_t^{(lab)}\right) \frac{1}{y_j} \right) \tag{5}$$

where $s_t^{(rec)}$ and $s_t^{(lab)}$ is the state/phone observed at frame t in the Viterbi and reference alignment, respectively, with T_n observations. $\delta(i,j)$ is the Kronecker delta. It is interesting to note that whenever there is a misalignment (different $s_t^{(rec)}$ and $s_t^{(lab)}$), then two outputs will always contribute to the gradient, in opposite directions. The output which agrees with the reference will contribute with a negative value and the wrong output with a positive value, telling the network to increase and decrease their corresponding values, respectively, according to the gradient descent algorithm. Figure 2 aims at illustrate the procedure considering the recognition of only four phones. If an error occurs (misclassification or misalignment) it will be given an indication to two outputs of the MLP.

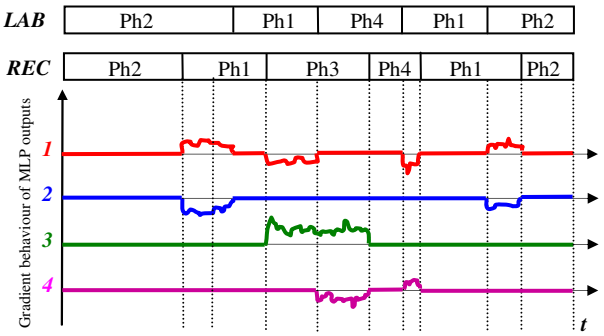

Figure 2: *Example of the gradient for each MLP output, in the presence of misclassifications or misalignments.*

Another interesting point is that the cost function based on Cross-Entropy, in the usual MLP training with targets, also has gradients inversely proportional to the outputs. However, a gradient term is computed for every frame, using the difference between the outputs and targets, which contrasts with the present global cost function that "blames" the MLP only when a misalignment occurs in the Viterbi and reference phone strings.

After computing the gradient of the cost function with respect to the MLP outputs we simply back-propagate gradients through the entire structure, all the way back to the first layer of the MLP. We used the resilient back propagation algorithm to accelerate the convergence to a solution.

2.3. Gradient with respect to the HMM parameters

The proposed hybrid MLP/HMM phone recognizer uses a Hidden Markov model to temporally align the speech signal, but instead of using a priori state-dependent observation probabilities defined by a Gaussian mixture, it uses a posteriori probabilities estimated by the MLP, keeping the overall HMM topology unchanged. In the hybrid system the output predictions of the MLP are interpreted as the *a posteriori* phone probabilities, $P(ph_i|\mathbf{x})$, with ph_i representing the i^{th} phone/state and **x** the feature observation vector.. In this way the only updatable HMM parameters are the state transitions, a_{ij}, which can be updated according to the following equation:

$$\begin{aligned}
\frac{\partial E}{\partial a_{ij}^{(m)}} &= \sum_{n=1}^{N_{BD}} \frac{\partial e^{(n)}}{\partial a_{ij}^{(m)}} \\
\frac{\partial e^{(n)}}{\partial a_{ij}^{(m)}} &= \sum_{k=1}^{N_W^{(rec)}} \delta(k,m) \left[\sum_{t=1}^{T_n} \left(\delta(s_{t-1}^{(rec)}, i)\delta(s_t^{(rec)}, j) \frac{1}{a_{ij}^{(m)}} \right) \right] \\
&\quad - \sum_{k=1}^{N_W^{(lab)}} \delta(k,m) \left[\sum_{t=1}^{T_n} \left(\delta(s_{t-1}^{(lab)}, i)\delta(s_t^{(lab)}, j) \frac{1}{a_{ij}^{(m)}} \right) \right]
\end{aligned} \tag{6}$$

3. Experimental Results

Phone recognition experiments were carried out using two different sets of speech material: English speech data from TIMIT database, [16] and European Portuguese speech data from TECNOVOZ database, [19].

Speech is analyzed every 10ms with a 25ms Hamming window. Thirty-nine parameters were used as standard input features of the MLPs representing 12 Mel Frequency Cepstral Coefficients (MFCCs), plus energy, and its first and second derivatives. The context window used is 170ms but only 9 frame features were used, one every other. The unused frame features are used in the next window analysis. The current frame is in the centre of the context window (temporal information of past and future is included), [20]. The *softmax* function was used as the activation function of the output layer so that the output values can be interpreted as posterior probabilities. The other hidden layer uses a sigmoid activation function. All the network weights and bias are adjusted using batch training with the resilient back-propagation (RP) algorithm [13] so as to minimize the minimum-cross-entropy error between network output and the target values.

The hidden Markov models used in the hybrid system were built for each phone (English and Portuguese separately) using HTK3.4, [15], in order to estimate the transition probabilities between states. Each phone was modeled by a three-state left-to-right HMM and each state was modeled by a single Gaussian model. In the hybrids MLP/HMM system, the a priori state likelihoods are replaced by the posterior probabilities given by the output predictions of the MLP. Each of the three states shares the same MLP output. We used HTK with some changes in order to replace the usual Gaussian mixture models with the outputs of the MLP. The performance was evaluated by means of Correctness (Corr) and Accuracy (Acc) using the HTK evaluation tool `HResults`.

3.1. TIMIT phone recognition task

When using TIMIT database, two single layer MLPs networks, with 1000 nodes, were trained for phone frame classification. In one, the last layer performs a 1-to-39 classification over the set of phones while in the other the last layer performs a 1-to-61 classification. Both training and testing were carried out using the TIMIT database [16]. In Baseline61 the original 61 phone set was used while in Baseline39 the train was made by means of the 39 phones proposed by Lee and Hon [17]. The training set consisted of all `si` and `sx` sentences of the original training set (3698 utterances) and the test set consisted of all `si` and `sx` sentences from the complete 168-speaker test set (1344 utterances). The targets derive from the phone boundaries provided by the TIMIT database. For evaluation purposes we collapsed the 61 TIMIT labels into the standard 39 phones, [17]. Table 1 shows the baseline results. Both systems achieved similar results. Baseline39 reached a *Correctness* rate of 72,79% and an *Accuracy* rate of 69,52% while for the Baseline 61 the corresponding rates are 72,46% and 69,60%.
In order to evaluate the training capabilities of the proposed discriminative training method, and also to achieve rapid convergence to a solution, the discriminative training method is implemented, starting from prior separately trained MLP and HMM systems, in a similar way to that reported in [2].
We will refer to the hybrid systems trained with the global discriminative training method as GDTM-MLP39/HMM and GDTM-MLP61/HMM. Results are presented in Table 1.

The results for GDTM indicate improvements both in Correctness and Accuracy. When using 39 phones, correctness rise up to 73.94, while when using 61 phones de improvement was of 1,37% (1,9 of relative improvement). With regard to accuracy the improvements are not so expressive (about 1% of relative) in both situations.

Table 1. *TIMIT Phone recognition results.*

System	Corr	Acc	(%) Relative Improvement	
			Corr	Acc
Baseline39	72.79	69.52	-	-
Baseline61	72.46	69.60	-	-
GDTM-MLP39/HMM	73.94	70.30	1.6	1.1
GDTM-MLP61/HMM	73.83	70.27	1.9	1.0

3.1.1. Comparison with other works

The results are not comparable with the ones posted in [18] and [5] because the authors of those works evaluated their systems by means of phone classification and not phone recognition, as we have done. But the results compare favorably with the findings presented by an ASAT (Automatic Speech Attribute Transcription) group in [3] and by Morris and Fosler-Lussier in [4]. These works have in common with the present work only the fact that they present results under the same conditions (same speech material and same recognition rates). The ASAT group, [3] uses confidence scores of phonetic attributes classes, coming from an MLP, an HMM and an SVM in a CRF for phone recognition. They point out a Corr rate of 73,39% and *Acc* rate of 69,52%. This value is similar to our Baseline results and below our GDTM-MLP/HMM results. Morris and Fosler-Lussier in [4] use phonological features provided by an ANN together with 61 class posteriors, provided by another ANN, also as input of a CRF. We did not yet reached their 71,49% Acc rate.

3.2. TECNOVOZ phone recognition task

TECNOVOZ is a European Portuguese speech database, [19] collected in 2007.. The collected speech includes commands and phonetically rich read sentences. The sentence utterances were divided into 20364 for the training set and 2262 for the testing set. To describe the Portuguese language 37 phones were used including a silence model and a short pause.

In the hybrid MLP/HMM system a single hidden layer MLP network, with 1000 nodes, was trained for frame-based phone classification. The last layer performs a 1-to-37 classification over the set of phones. The targets were obtained by forced alignment using the triphone model set described in [19].

Table 2 presents the baseline results. *Correctnes* reached 49.78% and *Accuracy* reached 45.59%. These results should be considered as preliminary because the number of the training iterations of the MLP was reduced and the used targets were not entirely verified. In fact, the triphone set used for forced alignment does not include all the triphones needed for this task/corpus. Thus, targets can be refined in order to accomplish the 37 phone recognition task so as to achieve better results.

Starting from this network and applying the proposed GDTM, *Correctnes* rise up to 55.43% (5.65% above) and *Accuracy* to 49.33% (3.74 above), representing 11.4% and 8.2% of relative improvement, respectively.

Besides the preliminary results, the same improvement trend observed on TIMIT task, was also verified with TECNOVOZ task, which indicates that the global training is useful.

Table 2. *TECNOVOZ Phone recognition results.*

System	Corr	Acc	(%) Relative Improvement	
			Corr	Acc
Baseline37	49.78	45.59	-	-
GDTM-MLP37/HMM	55.43	49.33	11.4	8.2

4. Conclusions

This paper describes a global discriminative training method (GDTM) applied to a hybrid MLP/HMM phone recognizer. The proposed method optimizes the network parameters as a function of the whole system. The MLP weights, which are usually updated according to the target values presented in the output layer, are now updated according to the misclassifications present in the output of the hybrid system. These misclassifications are computed comparing the output of the best Viterbi alignment with the reference alignment provided in the database. The gradient of the alignment errors are back propagated through the entire structure, all the way back to the first MLP layer. This results in a minimization of the classification error of the global hybrid system, and also maximizes the phone accuracy.

GDTM was tested using two databases: English TIMIT and Portuguese TECNOVOZ. In both tasks relative improvements in correctness and accuracy were achieved vis-à-vis the corresponding baselines.

5. Acknowledgements

Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant (SFRH/BD/27966/2006).

6. References

[1] E. Trentin, M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition”,. Neurocomputing. , vol. 37, pp. 91-126, March 2001.

[2] Bengio, Y., Mori, R., Flammia, G. and Kompe, R., “Global Optimization of a Neural Network-Hidden Markov Model Hybrid”, IEEE Transactions on Neural Networks, Vol. 3, No. 2, March 1992, pp 252-259.

[3] Bromberg, I., *et al*., "Detection-based ASR in the automatic speech attribute transcription project," in Proc. of Interspeech2007, pp. 1829-1832, August, 2007.

[4] Morris, J. and Fosler-Lussier, E., "Conditional Random Fields for Integrating Local Discriminative Classifiers," IEEE Transactions on Acoustics, Speech, and Language Processing, 16:3, pp 617-628, March 2008.

[5] Scanlon, P., Ellis, D. and Reilly, R., "Using Broad Phonetic Group Experts for Improved Speech Recognition", IEEE Transactions on Audio, Speech and Language Processing, vol.15 (3) , pp 803-812, March 2007.

[6] Wu, J., Huo, Q.: “An Environment-Compensated Minimum Classification Error Training Approach Based on Stochastic Vector Mapping”. IEEE Transactions on Audio, Speech & Language Processing 14(6): 2147-2155 (2006).

[7] Johansen, F.T., “Global discriminative modelling for automatic speech recognition”, Ph.D. thesis. , The Norwegian University of Science and Technology, Trondheim, Norway (1996).

[8] Droppo, J., Acero, A., “Joint Discriminative Front End and Back End Training for Improved Speech Recognition Accuracy”, in Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing. Toulouse, May, 2006.

[9] Woodland, P.C. and Povey. D. “Large scale discriminative training of hidden Markov models for speech recognition”. Computer Speech and Language, 16:25–47, 2002.

[10] Chou, W., Lee, C.-H., Juang, B.-H. and Soong, F.-K. “A minimum speech error rate pattern recognition approach to speech recognition,” Int. J. Pattern Recognition Artificial Intelligence, Special Issue on Speech Recognition for Different Languages, vol. 8, no. 1, pp. 5–31,1994.

[11] Povey, D. and Woodland, P., “Minimum phone error and i-smoothing for improved discriminative training,” in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2002, vol. 1, Orlando, FL, May 2002, pp. 105–108.

[12] Yu, Dong; Deng, Li, “Large-Margin Discriminative Training of Hidden Markov Models for Speech Recognition”, in Proc of the International Conference on Semantic Computing, 2007. Volume , Issue , 17-19 Sept. 2007 Page(s):429 – 438.

[13] Riis, S., Krogh, A. “Hidden Neural Networks: A Framework for HMM/NN Hybrids”, in Proc of ICASSP97 April, 1997.

[14] Riedmiller, M. and Braun, H. “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” in Proc. ICNN, San Francisco, CA, 1993, pp. 586–591.

[15] Young, S. *et al*, The HTK book. Revised for HTK version 3.4, Cambridge University Engineering Department, Cambridge, December 2006.

[16] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, n., DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST, 1990.

[17] Lee, K. and Hon, H., “Speaker-independent phone recognition using hidden Markov models”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37 (11), November 1989, pp. 1642-1648.

[18] Gunawardana, A., Mahajan, M., Acero, A., and Platt. J., “Hidden conditional random fields for phone classification,” in Proc. Interspeech, 2005, pp. 1117–1120.

[19] Lopes, J., Neves, C., Veiga, A., Maciel, A., Lopes, C., Perdigão, F., Sá, L., “Development of a Speech Recognizer with the Tecnovoz Database”, Propor 2008, International Conference on Computational Processing of Portuguese, Aveiro, Portugal.

[20] Lopes, C., Perdigão, F., "A Hierarchical Broad-Class Classification to Enhance Phone Recognition", 17th European Signal Processing Conference (EUSIPCO-2009), Glasgow, Scotland, August 2009.