A COMPARISON OF HYBRID HMM ARCHITECTURES USING GLOBAL DISCRIMINATIVE TRAINING

Finn Tore Johansen

Telenor Research and Development, N-2007 Kjeller, Norway
finn.johansen@fou.telenor.no

ABSTRACT

This paper presents a comparison of different model architectures for TIMIT phoneme recognition. The baseline is a conventional diagonal covariance Gaussian mixture HMM. This system is compared to two different hybrid MLP/HMMs, both adhering to the same restrictions regarding input context and output states as the Gaussian mixtures. All free parameters in the three systems are jointly optimised using the same global discriminative criterion. A Forward decoder, with total likelihood scoring, is used for recognition. While the global discriminative training method is found to improve the baseline HMM significantly, the differences between Gaussian and MLP-based architectures are small. The Gaussian mixture system however performs slightly better at the lowest complexity levels.

1. Introduction

Continuous density HMMs [1] and hybrid ANN/HMMs [2, 3] are two leading technologies for large vocabulary continuous speech recognition. In both these approaches, a Markov process is used to model the basic temporal nature of the speech signal, based on an input sequence of feature vectors. The actual architectures used for feature classification or observation modelling are however different. In continuous density HMMs, Gaussian mixture distributions are typically used to model state emission probabilities. In the hybrid systems, different ANN architectures (such as Multilayer Perceptrons (MLPs) [2, 4] or Recurrent Neural Networks (RNNs) [3] replace these Gaussians. In principle, Gaussian mixture HMM systems can also be regarded as ANN/HMM hybrids based on Radial Basis Function (RBF) networks [5, 6].

HMMs are conventionally trained by the Baum-Welch algorithm which maximises global model likelihood (ML). Discriminative training, usually with the MCE or MMI criterion has been shown to give improved performance, e.g. for connected digit recognition systems [7, 8]. ANN-based hybrids were originally trained to do single frame discrimination by the embedded Viterbi algorithm [9, 10]. Global discriminative training of ANN-based systems has however also been proposed [11, 12].

Since gradient-based global discriminative training can be done for general recogniser structures, it offers a unified framework for architecture comparisons. In this paper, the unified framework will be used to compare a conventional, context-independent Gaussian HMM recogniser structure to two slightly different MLP-based hybrid ANN/HMMs. All systems are identical with respect to feature analysis, Markov model topology and training and decoding methods. They only differ in the architectures used for acoustic observation modelling. It is our hope that such a comparison will be fair to both systems and ultimately provide us with the insight needed to find improved speech recogniser architectures.

The paper is organised as follows. Section 2 presents the criteria and algorithms used for training and testing. Section 3 presents the different model architectures; a baseline Gaussian HMM and two MLP/HMM hybrids. In Section 4, the experimental conditions and results are summarised, and the conclusion is given in Section 5.

2. Training and decoding

A Markov model output defines an a posteriori probability $P(w|x)$, for every observation sequence x and sentence transcription w. All model parameters can thus be optimised with the global discriminative criterion

$$E = -\sum_n \log P(w^n|x^n), \tag{1}$$

where n is the trainset index. This criterion is known as Conditional Maximum Likelihood (CML) as proposed by Brown in [13], and has also been referred to as the global MAP criterion [14, 12]. When the language model $P(w)$ is kept constant, CML/MAP is also equivalent to Maximum Mutual Information (MMI) training.

The criterion (1) is minimised by stochastic gradient search for all parameters θ,

$$\theta_{i+1} = \theta_i - \epsilon_i \frac{\partial E_i}{\partial \theta_i}, \quad (2)$$

where i is the update index, and E_i is the single sample error. A time-scaled inverse linear learning schedule

$$\epsilon_i = \frac{\epsilon_1}{\frac{i-1}{N_\epsilon} + 1}, \quad (3)$$

with $N_\epsilon = 100,000$ and $\epsilon_1 = 0.001$ is used.

Transitions, bigrams and Gaussian variances are updated in the logarithmic domain in order to maintain positive parameter values. To make stochastic gradient training work for Gaussian HMMs, it was found important to scale the Gaussian means on the corresponding standard deviations [15] before update.

No sum-to-one constraints are enforced for transitions or bigrams. This allows an effective state weighting and grammar scaling to take place. Without the sum-to-one constraints, even the conventional HMM architecture is not strictly a statistical model, but a recurrent network-based discriminator function [5, 16].

The decoder implements a search for the maximum likelihood sentence hypothesis, using full forward (or total likelihood) scoring,

$$w^* = \arg\max_w P(w|x). \quad (4)$$

This is different from Viterbi decoding criterion, which only considers a single state path from each sentence hypothesis. The time-synchronous search algorithm [17], with a maximum of five active hypotheses per state, has been found to give significantly better results than Viterbi decoding for global discriminative models [18].

3. Architectures

Three different model architectures are tried.

The *baseline* HMM architecture uses three states per context-independent phone model, in a forward connected topology with no skips. Each state contains a diagonal covariance Gaussian mixture density to model observation vectors. Twelve MFCC coefficients and a normalised log energy make up the basic feature vector. A five-frame linear regression is used to compute first order delta coefficients which are added to the static features.

Two different MLP/HMM hybrids are designed to meet the same constraints as the Gaussian layer in the HMM, namely an input span of five frames and three ouput states per phone.

In *MLP1/HMM*, the Gaussian mixture distributions, operating on 26-dimensional delta-extended features, are replaced by a single MLP with 26 input nodes and one output per HMM state. The MLP has one hidden layer, with a number of nodes selected so that the number of free parameters matches that of the Gaussian mixture models. Sigmoid nonlinearities are used in both the hidden and output nodes. Since the model is trained with a global discriminative criterion, there is no need to apply any prior scaling, as in [2].

The *MLP2/HMM* architecture is basically similar to MLP1/HMM. The only difference is that the 26-dimensional delta feature input is replaced by five consecutive 13-dimensional feature vectors. This leaves the MLP to decide how to model state-internal temporal feature dependencies. Because the input layer is now larger than in MLP1/HMM, the number of hidden nodes are reduced, so that the same overall complexity is achieved.

4. Experiments

The task selected for experiments is the standard TIMIT 39-class phone recognition task [19], with a full 3696-sentence trainset and 1344-sentence testset. Results are also presented on the smaller 192-sentence core testset.

One three-state model was used for each of the 39 phone classes. This makes a total of 117 states in the global recognition network. Phone bigrams were estimated from the training set and scaled by a factor of two, (squared probability values). This was done to match the diagonal covariance system reported in [20] as closely as possible.

Baseline Gaussian HMMs were trained by HTK [21], using twelve iterations of embedded Baum-Welch reestimation. Recognition results are given in Table 1. These are very similar to those reported in [20].

The ML-trained baseline models were then used as initial estimates in the discriminative optimisation, where bigrams, transitions, means and variances were jointly optimised during thirty epochs of stochastic gradient search. The resulting models are denoted CML/HMM in Table 1. For the single-mixture system, a fullbatch gradient optimisation, taken to complete convergence, was also performed [18]. This model is denoted CML*/HMM.

The MLP/HMM hybrids were trained by the same stochastic gradient algorithm as the CML/HMM models, using the ML model transitions and bigrams as initial Markov network parameters. The MLP weights were all initialised to random values. Initialising the MLPs to do frame classification had previously not been found to give improved performance [18]. Complete results are given in Table 2.

We see that the single-mixture CML*/HMM system has the overall best testset accuracy. It is significanlty better than even the 16 mixture baseline. The accuracy is comparable to the best results reported for context-independent HMMs [20,

Model	#Mix	#Param.	Trainset		Core testset		Full testset	
			%Corr	%Acc	%Corr	%Acc	%Corr	%Acc
Baseline	1	7917	58.29	54.18	56.34	52.61	57.20	53.06
Baseline	2	14118	62.38	57.89	60.68	56.80	61.22	56.97
Baseline	4	26376	66.41	62.08	63.81	59.58	64.50	60.18
Baseline	8	51324	69.20	65.74	65.93	62.37	67.07	63.17
Baseline	16	100932	72.07	69.15	67.89	64.67	69.02	65.49
CML/HMM	1	7917	71.07	68.39	67.14	64.21	68.66	65.47
CML/HMM	2	14118	73.69	70.83	69.23	65.82	71.01	67.43
CML/HMM	4	26376	77.10	73.69	70.70	66.14	72.01	67.47
CML*/HMM	1	7917	74.72	72.21	69.85	66.75	71.20	68.19

Table 1: Gaussian mixture HMM results

Model	MLP topology	#Param.	Trainset		Core testset		Full testset	
			%Corr	%Acc	%Corr	%Acc	%Corr	%Acc
MLP1/HMM	26-41-117	7854	69.99	67.28	65.92	62.84	66.92	63.96
MLP1/HMM	26-85-117	14190	71.61	69.07	68.61	65.86	69.71	66.85
MLP1/HMM	26-171-117	26574	74.53	72.23	69.15	65.99	70.84	67.88
MLP2/HMM	5x13-33-117	7989	69.15	66.47	66.15	63.01	66.71	63.68
MLP2/HMM	5x13-66-117	14028	72.45	69.85	67.80	64.49	69.20	66.11
MLP2/HMM	5x13-134-117	26472	75.32	72.75	69.45	66.21	70.70	67.51

Table 2: MLP/HMM hybrid results

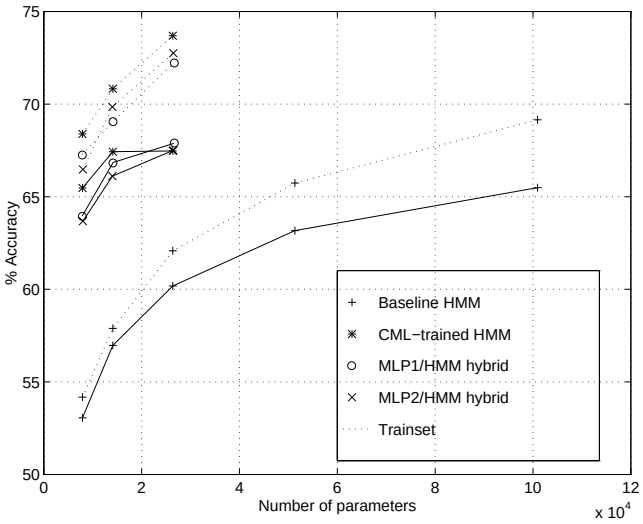

Figure 1: Comparison of recogniser accuracies

22], although the architecture is considerably simpler than these systems. The training complexity of the CML*/HMM was however too high for most practical applications, since more than 500 epochs were needed for convergence.

In Figure 1, recognition accuracies for the other models are presented as a function of the number of free parameters. All CML-trained systems are seen to be significantly better than the ML systems of comparable complexity. The difference between the three architectures is generally very small. The single-mixture Gaussian CML/HMM performes slightly better than the corresponding MLP/HMM hybrids, and the Gaussian models also seem to give slightly higher trainset accuracy.

5. Conclusion

In this paper, we have seen how global discriminative training combined with forward decoding can be used to compare different recogniser architectures. The method was applied to three different architectures for TIMIT phone recognition.

The applied training and decoding method improved the baseline HMM system significantly. The two MLP-based architectures seemed to be slightly worse than the Gaussian mixture system, although the performances of the three different architectures are very similar. This seems to indicate that all systems are limited by the same constraints, namely the five-frame input and HMM topology. With these constraints, the best generalisation performance could actually be reached with the simplest, single-mixture CML*/HMM.

In order to improve performance further, these constraints should probably be relaxed. This could e.g. be done by allowing more context input or by using context-dependent phone models. Input transformations on the features also represent an interesting possibility for the Gaussian system. The conventional three-state phone model topology, which has been selected on the basis of ML optimisation, should probably also be reconsidered within the framework of global discriminative optimisation.

6. REFERENCES

1. S. J. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech and Language*, vol. 8, pp. 369–383, Oct. 1994.
2. H. A. Bourlard and N. Morgan, *Connectionist speech recognition. A hybrid approach.* Boston: Kluwer Academic Publishers, 1994.
3. M. M. Hochberg, S. J. Renals, A. J. Robinson, and C. D. Cook, "Recent improvements to the ABBOT large vocabulary CSR system," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 69–72, 1995.
4. H. Franco, M. Cohen, N. Morgan, D. Rumelhart, and V. Abrash, "Context-dependent connectionist probability estimation in a hybrid hidden Markov model-neural net speech recognition system," *Computer Speech and Language*, vol. 8, pp. 211–222, 1994.
5. J. S. Bridle, "Alpha-nets: A recurrent 'neural' network architecture with a hidden Markov model interpretation," *Speech Communication*, vol. 9, pp. 83–92, Feb. 1990.
6. W. Reichl and G. Ruske, "A hybrid RBF-HMM system for continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 3335–3338, 1995.
7. Y. Normandin, R. Cardin, and R. De Mori, "High–performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, pp. 299–311, Apr. 1994.
8. W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training of inter-word context dependent acoustic models in speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pp. 439–442, Sept. 1994.
9. S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Acoust., Speech and Audio Processing*, vol. 2, pp. 161–174, Jan. 1994.
10. A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.
11. P. Haffner, "Connectionist speech recognition with a global MMI algorithm," in *Proc. Eur. Conf. Speech Comm. Tech. (EUROSPEECH)*, pp. 1929–1932, 1993.
12. H. Bourlard, Y. Konig, and N. Morgan, "REMAP: recursive estimation and maximization of a posteriori probabilities — application to transition-based connectionist speech recognition," Tech. Rep. TR-94-064, ICSI, Mar. 1995.
13. P. F. Brown, *The acoustic modeling problem in automatic speech recognition.* PhD thesis, Carnegie Mellon University, May 1987.
14. N. Morgan and H. A. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proc. IEEE*, vol. 83, pp. 742–770, May 1995.
15. W. Chou, B. H. Juang, and C. H. Lee, "Segmental GPD training of HMM based speech recognizer," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. I–473–476, 1992.
16. L. Niles, "TIMIT phoneme recognition using an HMM-derived recurrent neural network," in *Proc. Eur. Conf. Speech Comm. Tech. (EUROSPEECH)*, pp. 559–562, 1991.
17. R. Schwartz and Y.-L. Chow, "The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. 81–84, 1990.
18. F. T. Johansen, *Global discriminative modelling for automatic speech recognition.* PhD thesis, Norwegian University of Science and Technology, May 1996.
19. K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1641–1648, Nov. 1989.
20. S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *Proc. Int. Conf. Acoust., Speech, Sign. Proc. (ICASSP)*, pp. II–491–494, Apr. 1993.
21. S. J. Young, *HTK: Hidden Markov Model Toolkit V1.4.* CUED Speech Group, Sept. 1992.
22. R. De Mori, M. Galler, and F. Brugnara, "Search and learning strategies for improving hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 107–121, Apr. 1995.