DISCRIMINATIVE MIXTURE WEIGHT ESTIMATION FOR LARGE GAUSSIAN MIXTURE MODELS

Françoise Beaufays, Mitchel Weintraub, Yochai Konig

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA.

ABSTRACT

This paper describes a new approach to acoustic modeling for large vocabulary continuous speech recognition (LVCSR) systems. Each phone is modeled with a large Gaussian mixture model (GMM) whose context-dependent mixture weights are estimated with a sentence-level discriminative training criterion. The estimation problem is casted in a neural network framework, which enables the incorporation of the appropriate constraints on the mixture weight vectors, and allows a straight-forward training procedure, based on steepest descent.

Experiments conducted on the Callhome-English and Switchboard databases show a significant improvement of the acoustic model performance, and a somewhat lesser improvement with the combined acoustic and language models.

1. INTRODUCTION

Many factors contribute to the relatively high error rates observed in LVCSR systems (*e.g.* diversity of speaking styles, pronunciation variants, variable degrees of articulation, noises, channel effects). By enlarging the set of possible acoustic realizations for each phone or phone state, these factors cause the acoustic models to have broad overlapping distributions, which in turn increases the chances of misrecognition. For this same reason of acoustic diversity, the amount of training data and the efficiency with which it is used are critical factors in the performance of LVCSR speech recognizers.

The modeling technique described in this paper addresses these two issues: it makes efficient use of the training data by allowing a maximum of data sharing between different states of each phone, and it attempts to limit the overlap between phone models by training them in a discriminative fashion.

In the past few years, the trend in acoustic modeling has been to define triphone clusters, and model each cluster with a relatively small number of Gaussians. Such clusters were derived through acoustic-based agglomerative clustering of triphone states (*e.g.* [1]), or with the help of binary decision trees based on contextual linguistic features (*e.g.* [2]). Triphone-cluster modeling offered a more detailed modeling of co-articulation effects than previously used phonetically tied mixture (PTM) models, and improved recognition accuracy [1, 2].

Recently, however, large PTM systems were shown to offer an attractive alternative to triphone-cluster models [3]. N-best lists rescoring experiments on the Callhome-Spanish database showed that context-independent phone GMMs could lead to the same word error rates as context-dependent decision-tree models, provided that the number of Gaussians in each phone model was sufficient, *i.e.* comparable to the total number of Gaussians in all the leaf nodes of the corresponding trees [3].

The main advantage of large PTM systems over small triphone-cluster models is the sharing of all the data aligned to a specific phone to train its model. This advantage may however be offset by the increased acoustic space coverage of each model and hence by the increased possibility of overlap between model distributions.

In this paper, we focuss on improving the discriminative power of large PTM systems by re-estimating the mixture weights of the phone GMMs for different contexts, using a discriminative optimization criterion. Various approaches to discriminative training of acoustic models have been proposed in the literature, *e.g.* [4, 5, 6, 7]. Because we wish to use the discriminative PTM models to rescore N-best lists of hypotheses, and because we believe that a global (sentence-level) training criterion will be more closely related to the recognition metric (here, the word error rate (WER)), we propose to train the models to maximize the average log-posterior probability of the correct transcriptions of the training sentences.

2. DISCRIMINATIVE MIXTURE WEIGHT ESTIMATION

Each phone in the context-dependent (CD) PTM system is modeled with a Gaussian mixture model, according to

$$p(\mathbf{x}_k|\varphi,i) = \sum_{g=1}^{N_\varphi} P_g^i N_g(\mathbf{x}_k), \qquad (1)$$

where φ indicates the phone being modeled, i refers to a specific context realization or triphone cluster of φ, and P_g^i and $N_g(\mathbf{x}_k)$ represent, respectively, the g^{th} context-dependent mixture weight in cluster i, and the g^{th} context-independent Gaussian distribution evaluated for the observation $\mathbf{x}_k$. In our implementation, triphone clusters are generated with linguistically-driven decision trees (DTs), so that the index i indicates the i^{th} leaf node of the DT corresponding to phone φ. (To be more precise, three DTs are built for each phone (one per state), and i indexes the leaf nodes across all three DTs.)

Given a set of data clusters for each phone, the problem is to estimate the corresponding mixture weight distributions. Because the mixture weights must satisfy $0 \leq P_g^i \leq 1$ and $\sum_g P_g^i = 1$, this is a constrained optimization problem. The approach we follow here is to cast this estimation problem in a neural network framework, as depicted in Fig. 1.

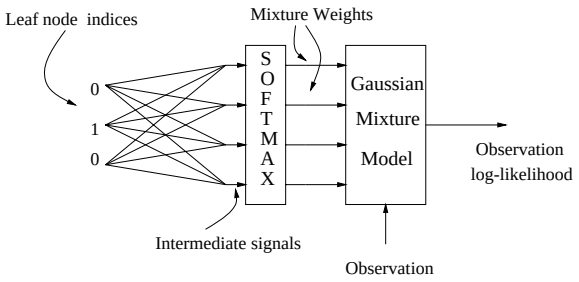

Figure 1: Neural network representation of the mixture weight estimation procedure.

The set of mixture weight vectors for all the contexts of a phone can be seen as a two-layer feedforward neural network (NNet) whose inputs are all zero, besides the one that corresponds to the cluster index of the current observation, which is set to one. The outputs of the NNet are passed through a softmax nonlinearity [8] to ensure that the above constraints on the mixture weights are satisfied.

According to the usual NNet methodology, an optimization criterion can then be defined, and the mixture weights can be iteratively updated, following a steepest descent approach.

2.1. Discriminative Training of the Mixture Weights

As mentioned previously, we wish the training criterion to be discriminative, closely related to the recognition error metric (the WER), expressible in an N-best list framework, and – to make a steepest descent approach feasible – continuous in the parameters to optimize, *i.e.* the mixture weights. One such criterion is the average log-posterior probability of the correct transcriptions of the training sentences,

$$\xi = \frac{1}{N_s} \sum_{s=1}^{N_s} \log P(W_c^s \mid \mathcal{X}_s) \quad (2)$$

$$P(W_c^s \mid \mathcal{X}_s) = \frac{p(\mathcal{X}_s, W_c^s)}{p(\mathcal{X}_s, W_c^s) + \sum_{h=1}^{N_h} p(\mathcal{X}_s, W_h^s)}, \quad (3)$$

where $\mathcal{X}_s = [\mathbf{x}_1, ..., \mathbf{x}_k, ...\mathbf{x}_K]$ denotes the sequence of acoustic observations for sentence s, W_c^s and W_h^s denote the word sequences in the correct transcription and in the h^{th} hypothesis of sentence s, $P(W_c^s \mid \mathcal{X}_s)$ denotes the posterior probability of the correct transcription given the acoustic observations, and N_s and N_h denote, respectively, the number of training sentences and the N-best list depth.

This criterion is similar to the N-best list implementation of the maximum mutual information criterion [5, 9], except that we include the joint probability of the observations and the correct word sequence in the denominator. This modification enables an intuitive interpretation of the mixture weight update formula.

The joint probabilities in Eq. 3 can be expanded into products of language and acoustic model probabilities (corrected by the language model weight, λ):

$$p(W_{c/h}^s, \mathcal{X}_s) = p_{LM}(W_{c/h}^s)\, p_{AM}(\mathcal{X}_s \mid W_{c/h}^s)^{1/\lambda}. \quad (4)$$

The neural network weights are adapted proportionally to the sentence-level instantaneous gradient of ξ,

$$\hat{\nabla}_\Theta\, \xi = \nabla_\Theta \log P(W_c^s \mid \mathcal{X}_s) \quad (5)$$

$$= \sum_h P_s(h) \Big[\nabla \log p_{AM}(c) - \nabla \log p_{AM}(h) \Big] \quad (6)$$

where Θ denotes the set of all the mixture weights, and where, with the independence assumption, $\log p_{AM}(.)$ can be rewritten as a sum of frame log-likelihoods.

Eq. 6 can be interpreted as follows. First, the weighted sum over the hypotheses gives more importance to the hypotheses whose posterior probabilities, $P_s(h)$, are larger. Second, for a given reference-hypothesis pair, (c, h), the mixture weights are adjusted only for the frames $\mathbf{x}_k$ for which the reference and hypothesis strings do not coincide ($p_{AM}(\mathbf{x}_k \mid c) \neq p_{AM}(\mathbf{x}_k \mid h)$). In that case, positive training is given to the correct model (c) and negative training is given to the erroneously hypothesized model (h). This is overall an intuitively satisfactory behavior. In practice, however, we slightly modify this update procedure: Eq. 6 implies that a weight update would occur also for the frames and model pairs (c, h) whose phone labels agree but whose triphone labels are different. Because we don't want to discriminate between allophones of the same phone, we instead perform the weight update only when the phone labels of the reference and hypothesis strings are different.

Note that, alternatively to a log-posterior criterion, we could have optimized the (linear) posterior probability of the correct sentences. We chose not to do this because, since $\nabla P(W_c^s) = P(W_c^s) \nabla \log P(W_c^s)$, it would have introduced a multiplicative term $P(W_c^s \mid \mathcal{X}_s)$ in Eq. 6, with the effect of giving less training to the sentences with small probability of being correct, which we thouhgt was intuitively undesirable.

To summarize the algorithm, the set of mixture weights Θ_n at time n is updated according to

$$\Theta_{n+1} = \Theta_n + \Delta\Theta_n \quad (7)$$

$$\Delta\Theta_n = \mu\, \hat{\nabla}_{\Theta_n} \xi, \quad (8)$$

where μ is a constant that governs the learning rate, $\hat{\nabla}_{\Theta_n} \xi$ is given by Eq. 6, and where the gradients of the frame log-likelihoods are backpropagated through the GMMs and the NNets output nonlinearities to update the neural network parameters.

3. RECOGNITION EXPERIMENTS

3.1. Baseline System and Database

The baseline system for this work is a speaker-independent speech recognition system based on continuous-density, genonic hidden Markov models [1]. It uses a multi-pass recog-

nition strategy, with a vocabulary of 33,275 words, non-cross word acoustic models and a trigram interpolated language model. Its training data consists of a mix of Callhome-English and Switchboard conversations. The total number of training sentences was 121K male sentences and 149K female sentences. The test data consists of the NIST-defined Eval'97 test set, which contains 1.7K male and 2.8K female sentences. The baseline system was used to generate N-best lists for all the training and testing data.

3.2. Experiment Description

As a first step, we trained two sets of context-dependent DTs (one per gender), to be used as a triphone clustering tool. All the training data was used to train the DTs. A stopping criterion of 1280 frames per leaf node was imposed to ensure that each cluster contains enough data.

We then trained two sets of moderate size context-independent (CI) PTM models (up to circa 450 Gaussians per phone). The exact number of Gaussians in each phone model was made equal to the number of leaf nodes in the corresponding DTs, that is about $1/16^{th}$ or $1/32^{th}$ of the total number of Gaussians we would have used in DT-based triphone models.

Context-dependent discriminative mixture weights were then estimated by looping over the randomized training sentences, aligning the top-N hypotheses with the CI models, computing the posterior probabilities of each hypothesis, adapting the mixture weights according to Eqs. 6-8, and repeating the process for several training epochs. To reduce the training time, we limited the N-best list depth to 5 hypotheses, and trained the models with only 10K (male) and 6K (female) sentences chosen at random from the training set (half Switchboard, half Callhome). The convergence process was monitored by computing the performance of the models on Eval'97 after each training epoch. No learning rate scheduling was used, so that the test set performance was used solely to decide when to stop training the weights, not to regulate the gradient step sizes.

WERs on the test sets were computed based on acoustic scores (log-likelihoods of the hypotheses) as well as based on combined acoustic and language model scores (Eq. 4) (later refered to as "AC" and "AC+LM", respectively). The test sentences were rescored with three different N-best list depths: 20, 10, and 5 hypotheses. After increasing the vocabulary size in the baseline system, new sets of N-best lists were generated for the test data, and the performance of the converged models was re-evaluated. On the male data, with the improved N-best lists, the average log-posterior of the correct sentences improved from -7.21 to -4.93 as a result of the discriminative training, and the average (linear) posterior probability increased from 0.19 to 0.30 (chance would be $1/6 = 0.166$, since the posterior probabilities are computed over 5 hypothesis and one reference strings). Similar numbers were obtained in the other three cases.

Tables 1-4 report the WERs with the CI and discriminative CD models. They show that, for both genders and for both sets of N-best lists, the discriminative training of the PTM mixture weights significantly improved the performance of the models (from 1.3 and 3.4 % relative depending on the experiment, with 5 hypotheses). However, the improvement brought by the language model somewhat outweighted the acoustic gain, especially for the male models. This came as a surprise since the language model is included in the training criterion, via the computation of the posterior probabilities of the hypotheses.

# hyps	GMMs		NNets	
	AC	AC + LM	AC	AC + LM
20	66.09	60.52	63.82	59.72
10	64.60	59.86	62.54	59.10
5	63.07	59.65	61.36	58.87

Table 1: Rescoring WER with GMMs and NNet models. Original N-best lists. Female test set Eval'97, 2.8K CH+SWB sentences.

# hyps	GMMs		NNets	
	AC	AC + LM	AC	AC + LM
20	66.83	61.31	65.00	61.72
10	65.44	60.77	63.19	60.68
5	63.68	60.44	62.07	60.35

Table 2: Rescoring WER with GMMs and NNet models. Original N-best lists. Male test set Eval'97, 1.7K CH+SWB sentences.

# hyps	GMMs		NNets	
	AC	AC + LM	AC	AC + LM
20	57.59	53.30	55.70	52.41
10	56.41	52.67	54.52	51.84
5	55.17	52.24	53.29	51.58

Table 3: Rescoring WER with GMMs and NNet models. Improved N-best lists. Female test set Eval'97, 2.8K CH+SWB sentences.

4. CONCLUSION

Motivated by the intuition that the high error rates observed with LVCSR databases are caused in great part by large overlaps between phone distributions, we proposed an acoustic modeling architecture in which each phone is modeled with a large Gaussian mixture model whose mixture weights are estimated in a context-dependent fashion by optimizing a sentence-level discriminative criterion between phones. The estimation process was casted in a neural network framework, and the mixture weights were optimized using a steepest descent approach.

Experiments with relatively small size models showed a significant improvement of the acoustic models, although the error-rate reduction brought by the language model

# hyps	GMMs		NNets	
	AC	AC + LM	AC	AC + LM
20	60.43	55.60	59.00	56.06
10	58.83	55.20	57.71	55.54
5	57.37	54.76	56.64	54.68

Table 4: Rescoring WER with GMMs and NNet models. Improved N-best lists. Male test set Eval'97, 1.7K CH+SWB sentences.

somewhat outweighted this gain. This point remains to be investigated. It may be due to the fact that the language model is given too much relative importance by being incorporated in the acoustic model update as well as being added to the acoustic model scores in the traditional way (Eq. 4).

Current experiments include training large PTM systems (up to 2000 Gaussians per phone), with N-best lists depths of 10 and 100 hypotheses.

5. REFERENCES

[1] V. V. Digalakis and P. Monaco and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers", *IEEE Trans. Speech, Audio Processing*, vol.4(4), July 1996, pp. 281-289.

[2] S. J. Young and J. J. Odell and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling", *Proc. Human Language Tech. Workshop*, March 1994, pp. 307-312.

[3] F. Beaufays, M. Weintraub, and Y. Konig, "DYNAMO: An Algorithm for Dynamic Acoustic Modeling", *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop.*

[4] P. Brown, "The acoustic modeling problem in automatic speech recognition", Ph.D. thesis, CS dept. CMU, 1987.

[5] L. R. Bahl and P. F. Brown and P. V. de Souza and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", *ICASSP* 1986, pp. 49-52.

[6] S. Katagiri and C.-H. Lee and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method", *Proc. Workshop on Neural Networks for Signal Proc.*, 1991, pp. 299-308.

[7] H. Bourlard and Y. Konig and N. Morgan, "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities", Tech. Rep. TR-94-064, ICSI, Berkeley, March 1995.

[8] J. S. Bridle, "Training Stochastic Model Recognition Algorithms as Networks can lead to Maximum Mutual Information Estimation of Parameters", in *Advances in Neural Information Processing Systems*, vol.2, Ed. D. S. Touretzky, Morgan Kaufmann, 1990.

[9] Y. L. Chow, "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm", *ICASSP* 1990, pp. 701-704.