DISCRIMINATIVE TRAINING OF HIERARCHICAL ACOUSTIC MODELS FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Hung-An Chang and James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts, 02139, USA
{hung_an, glass}@csail.mit.edu

ABSTRACT

In this paper we propose discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition tasks. After presenting our hierarchical modeling framework, we describe how the models can be generated with either Minimum Classification Error or large-margin training. Experiments on a large vocabulary lecture transcription task show that the hierarchical model can yield more than 1.0% absolute word error rate reduction over non-hierarchical models for both kinds of discriminative training.

Index Terms— hierarchical acoustic modeling, discriminative training, LVCSR

1. INTRODUCTION

There has been much effort devoted to improving the acoustic modeling component of automatic speech recognition (ASR) systems for large-vocabulary continuous-speech recognition (LVCSR) tasks. In recent years, discriminative training methods have demonstrated considerable success; these methods seek to reduce model confusion according to an objective function that is related to the error rate. Several kinds of discriminative training criteria such as Maximum Mutual Information, Minimum Phone Error, and Minimum Classification Error (MCE) have been proposed in the literature and have been shown to be effective in reducing the word error rate (WER) on a variety of LVCSR tasks.

Another approach to improve the acoustic modeling component is to utilize a more flexible model structure by constructing a hierarchical tree for the models. A hierarchy partitions the classification problem into smaller sub-problems that may be easier to model. In addition, a hierarchical model may be more robust, as non-terminal nodes in the hierarchy are less likely to suffer over-fitting since there are more training exemplars. Hierarchical models have been shown to be effective for the task of phonetic classification. In [1], the hierarchy was used to combine different acoustic measurements to improve the classification performance; in [2], a manually constructed hierarchical model was integrated with large-margin training [3] and shown to have better classification performance using fewer parameters than a conventional flat acoustic model. Since hierarchical modeling has shown significant benefits for phonetic classification tasks, we were interested to explore whether these methods could be applied to LVCSR tasks as well.

In this paper, we propose a hierarchical acoustic modeling scheme that can be trained using discriminative methods for LVCSR tasks. A model hierarchy is constructed by first using a top-down divisive clustering procedure to create a decision tree; hierarchical layers are then selected by using different stopping criterion to traverse the decision tree. Parameters in the hierarchical model are then learned using a discriminative training method; in this paper we explore both MCE training [4] and large-margin training [3]. The performance of the proposed modeling scheme is evaluated on a large-vocabulary lecture transcription task [5].

The organization of the paper is as follows. In section 2, the method for constructing the hierarchical acoustic models is presented, and the way to score the models is derived. Section 3 briefly introduces the discriminative training algorithms used in the experiments and we describe how to combine hierarchical modeling with discriminative training methods. Experimental results on the lecture task are reported in section 4, followed by some concluding remarks.

2. HIERARCHICAL GAUSSIAN MIXTURE MODELS

2.1. Model Hierarchy Construction

Top-down decision tree clustering [6] is a sequential algorithm that is often used for context-dependent acoustic modeling on LVCSR tasks. The algorithm uses a tree structure to represent the current status of clustering, and each node in the tree represents a set of acoustic contexts and their corresponding training data. The algorithm begins with a single root node, and at each step of the algorithm, one leaf node of the tree is split into two according to a context related ques-

This work is supported by Taiwan Merit Scholarship (Number NSC-095-SAF-I-564-040-TMS), by the T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan, and by Nippon Telegraph and Telephone Corp., Japan.

tion. In general, the node and the question are chosen such that the clustering objective such as the data log-likelihood can be optimized at each step. The algorithm continues until some stopping criterion such as the maximum number of leaf nodes is reached.

A hierarchical model can be constructed from a decision tree by running the clustering algorithm multiple times using different stopping criteria. Fig. 1 illustrates how to construct a 2-level hierarchy. The lightly-shaded nodes A, B, and C on the left side of Fig. 1 are produced by the first run of the clustering; by using a looser stopping criterion, the algorithm can continue splitting the nodes, and the dark-shaded nodes D, E, F, G, and H can be generated. A model hierarchy can thus be constructed by setting the lightly shaded nodes to be the parent nodes of the dark-shaded nodes, as illustrated in the right side of Fig. 1.

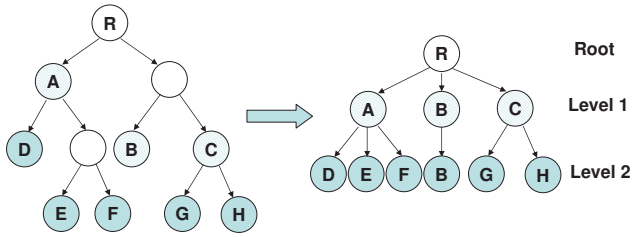


Fig. 1. Example of constructing model hierarchy by running top-down decision tree clustering multiple times with different stopping criterion.

2.2. Model Scoring

The leaf nodes of a hierarchical model represent the set of its output acoustic labels; during decoding, the speech recognizer requires acoustic scores for these output labels. For each non-root node c in the hierarchy, a set of Gaussian mixture model (GMM) parameters can be used to model the training data corresponding to c. As a result, the log-likelihood of a feature vector $\mathbf{x}$ with respect to c can be computed by

$$l_{\boldsymbol{\lambda}}(\mathbf{x}, c) = \log(\sum_{m=1}^{M_c} w_{cm}\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_{cm}, \boldsymbol{\sigma}_{cm})), \quad (1)$$

where m is the index of mixture component, M_c is the total number of mixture components of c, w_{cm} is the mixture weight of the m^{th} component of c, and $\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_{cm}, \boldsymbol{\sigma}_{cm})$ is the multivariate Gaussian density function of $\mathbf{x}$ with respect to mean vector $\boldsymbol{\mu}_{cm}$ and standard deviation $\boldsymbol{\sigma}_{cm}$. For each output label p in a K level hierarchy, let $\mathcal{A}_k(p)$ denote the parent of p at the k^{th} level of the hierarchy ($\mathcal{A}_K(p) = p$). The output acoustic score of p can be computed with a weighted average over the log-likelihoods of its parents:

$$l^H_{\boldsymbol{\lambda}}(\mathbf{x}, p) = \sum_{k=1}^{K} \omega_k l_{\boldsymbol{\lambda}}(\mathbf{x}, \mathcal{A}_k(p)), \quad (2)$$

where ω_k denotes the relative weight of the k^{th} level.

3. DISCRIMINATIVE TRAINING

The following sections describe the two discriminative training algorithms, Minimum Classification Error (MCE) training [4] and Large-Margin (LM) training [3], used in our experiments in more detail.

3.1. MCE Training

MCE training seeks to minimize the number of incorrectly recognized utterances in the training set by increasing the difference between the log-likelihood score of the correct transcription and that of an incorrect hypothesis. Let $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ denote the log-likelihood scores of the hypothesis $\mathbf{S}$ given the feature vectors $\mathbf{X}_n$ of the n^{th} training utterance. Note that $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ can be computed by summing the acoustic model scores of all p as in Equation (2) related to $\mathbf{S}$ with corresponding pronunciation and language model scores.

The loss function of MCE training is often smoothed by a differentiable sigmoid function $\ell(d) = \frac{1}{1+\exp(-\zeta d)}$, where ζ is a positive constant determining the sharpness of the function. In our experiments, an N-Best version of MCE training is implemented; that is, given a training set of N utterances, the loss function can be expressed by

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \ell[-L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y_n}) + \\ \log([\tfrac{1}{C}\sum_{\mathbf{S}\in\boldsymbol{\mathcal{S}}_n} \exp(\eta L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}))]^{\frac{1}{\eta}})],\end{aligned} \quad (3)$$

where $\mathbf{Y_n}$ denotes the transcription for the n^{th} utterance, $\boldsymbol{\mathcal{S}}_n$ denotes the set of N-Best incorrect hypotheses, C denotes the size of the N-Best list, and η is set to 1.0 in our experiments. Given the loss function, the model parameters are updated using a gradient-based method call *QuickProp*[4].

3.2. Large-Margin Training

Large-margin (LM) training proposed by Sha and Saul [3] has been shown to have good performance for the task of phonetic recognition. In LM training all the likelihood scores $L_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ are transformed into a set of distance metrics $D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$. For each incorrect hypothesis $\mathbf{S}$, the LM constraint requires the distance metric $D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S})$ to be greater than $D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y}_n)$ by a margin proportional to the Hamming distance $H(\mathbf{Y}_n, \mathbf{S})$ between the hypothesis and the transcription. For each violation of the constraint, the difference

$$D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y}_n) - D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}) + \gamma H(\mathbf{Y}_n, \mathbf{S}) \quad (4)$$

is considered to be a loss, where γ is a positive constant. By relaxing the above constraints for all S in the N-Best list of each training utterance, and summing over the loss across all utterances, the loss function for LM training can be expressed by

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \alpha_n[D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y_n}) + \\ \log([\tfrac{1}{C}\sum_{\mathbf{S}\in\boldsymbol{\mathcal{S}}_n} \exp(\gamma H(\mathbf{Y}_n, \mathbf{S}) - D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{S}))])]_+,\end{aligned} \quad (5)$$

where α_n is an utterance weight used to prevent outliers. The values of α_n and γ need to be specified before the training, and some heuristic approaches for selecting appropriate values can be found in [7]. Note that if we transform the GMM parameters appropriately [3] and further relax the distance metric $D_{\boldsymbol{\lambda}}(\mathbf{X}_n, \mathbf{Y}_n)$, the loss function can be convex to the transformed parameters. In our implementation, given the loss function of the training data, the model parameters can also be updated by the *Quickprop* algorithm.

3.3. Discriminative Training with Hierarchical Model

Here we show how to combine a hierarchical acoustic model with discriminative training. Given the loss function $\mathcal{L}$, the gradient $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}}$ can be decomposed into a combination of the gradients of the individual acoustic scores, as would be done for discriminative training of non-hierarchical models:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \sum_{n=1}^{N} \sum_{\mathbf{x} \in \mathbf{X}_n} \sum_{p} \frac{\partial \mathcal{L}}{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, p)} \frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x}, p)}{\partial \boldsymbol{\lambda}}, \qquad (6)$$

where $a_{\boldsymbol{\lambda}}(\mathbf{x}, p)$ denotes the acoustic model score for feature $\mathbf{x}$ and output label p. Since the acoustic score of a hierarchical model can be computed by Equation (2), $\frac{\partial a_{\boldsymbol{\lambda}}(\mathbf{x},p)}{\partial \boldsymbol{\lambda}}$ can be decomposed into $\sum_{k=1}^{K} \omega_k \frac{\partial l_{\boldsymbol{\lambda}}(\mathbf{x}, \mathcal{A}_k(p))}{\partial \boldsymbol{\lambda}}$. As a result, the training on a hierarchical model can be reduced to first computing the gradients with respect to all acoustic scores as would be done in the training for a non-hierarchical model, and then distribute the contribution of the gradient into different levels according to ω_k.

4. EXPERIMENTS

4.1. MIT Lecture Corpus

The MIT Lecture Corpus contains audio recordings and manual transcriptions for approximately 300 hours of MIT lectures from eight different courses, and nearly 100 MITWorld seminars given on a variety of topics [5]. The recordings were manually transcribed in a way that disfluencies such as filled pauses and false starts are labeled.

Among the lectures in the corpus, a 119-hour training set that includes 7 lectures from 4 courses and 99 lectures from 4 years of MITWorld lectures covering a variety of topics is selected for the acoustic model training. Two held-out MITWorld lectures (about 2 hours) are used for model development such as deciding when to stop the discriminative training. The test lectures are composed of 8 lectures from 4 different classes with roughly 8 hours of audio data and 7.2K words. Note that there is no speaker overlap between the three sets of lectures. More details of the lectures can be found in [7].

4.2. SUMMIT Recognizer

Instead of extracting feature vectors at a constant frame-rate as in conventional Hidden Markov Model (HMM) speech recognizers, the SUMMIT landmark-based speech recognizer [8] first computes a set of perceptually important time points as landmarks based on an acoustic difference measure, and extracts a feature vector around each landmark. The landmark features are computed by concatenating the average values of 14 Mel-Frequency Cepstrum Coefficients in 8 telescoping regions around each landmark (total 112 dimensions), and are reduced (and whitened) to 50 dimensions by Principal Component Analysis. The acoustic landmarks are represented by a set of diphones labels to model the left and right contexts of the landmarks. The diphones are clustered using top-down decision tree clustering and are modeled by a set of GMM parameters. All the other constraints for LVCSR, including pronunciation rules, lexicon, and language models are represented by Finite-State Transducers (FSTs), and speech recognition is conducted by performing path search in the FST [9].

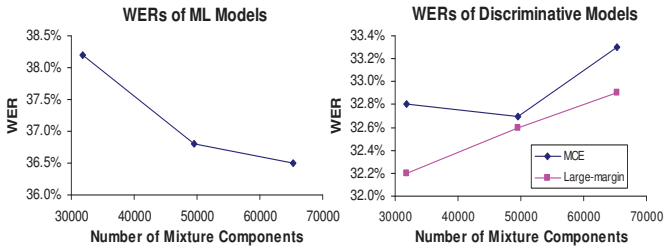
4.3. Baseline Models

There are total of 5,549 diphone labels used by SUMMIT. For the baseline models, the diphones were clustered into 1,871 diphone classes using a top-down decision tree clustering algorithm; initial GMMs for the diphones were created from the training data using the Maximum Likelihood (ML) criterion. To see how the number of parameters affects the model performance, three ML models were trained with different numbers of mixture components. Figure 2(a) summarizes the performance of the ML models. As illustrated in Figure 2, the WERs were reduced as the number of mixtures was increased, showing that increasing the number of parameters provides better fitting under the ML training criterion.

MCE and LM training were applied to the ML models, and two discriminative models were generated for each initial ML model. Figure 2(b) summarizes the WERs of the discriminatively trained models. Although discriminative training always reduced the WERs, the improvements were not as effective as the number of mixture component increased. Comparing the MCE loss function of the models on both the training and test data, we found that although the training loss decreases as the number of mixture components increases, the test loss increases. This fact suggests that discriminative training has made the model over-fit the training data as the number of parameters increases.

4.4. Hierarchical Models

By using a different stopping criterion for the top-down decision tree clustering, another diphone set with 773 diphone classes was generated. Although we didn't show the results,



(a) ML Models (b) Discriminative Models

Fig. 2. Test set WERs of non-hierarchical models. The WERs of ML models reduce as the number of mixture components increases, but the gain is not translated after discriminative training.

an MCE trained model with these classes show the same WER trend found in Fig. 2(b) with about 0.5% higher error.

As described in section 2, the 773-diphone cluster and the original 1871-diphone cluster can form a 2-level model hierarchy. A ML model with about 22K mixture components was trained for the 773-diphone cluster, and the model was combined with the baseline ML model of about 32K mixture components to form a hierarchical model with a total of approximately 54K mixture components. The ω_1 (for the 733-diphone cluster) and ω_2 in Equation (2) were set to 0.4 and 0.6 respectively.

Table 1 summarizes the WERs of the hierarchical model before and after discriminative training. Before discriminative training, the hierarchical model has a higher WER on the test set than the non-hierarchical model with a similar number of parameters. However, for both discriminative training methods, the hierarchical model yields more than 1.0% WER reduction over the best non-hierarchical model, i.e. 32.7% for MCE training and 32.2% for LM training from Fig. 2(b). The McNemar significance test [10] shows that the improvement of the hierarchical model over the non-hierarchical ones is statistically significant ($p < 0.001$). To understand why the hierarchical model generalize better, we measured the log-likelihood differences between the ML models and the discriminatively trained models. The statistics of log-likelihood differences show that the hierarchy prevents large decreases in the log-likelihood and can potentially make the model be more resilient to over-fitting.

	ML	MCE	LM
WER	37.4%	31.2%	**31.0%**

Table 1. Test set WER of the hierarchical model before and after discriminative training.

5. CONCLUSION AND FUTURE WORK

In this paper, we have described how to construct a hierarchical model for LVCSR task using top-down decision tree clustering, and how to combine a hierarchical acoustic model with discriminative training. In experiments using the MIT Lecture Corpus, the proposed hierarchical model yielded over 1.0% absolute WER reduction over the best-performing non-hierarchical model. In the future, we plan to further improve the model by applying a more flexible hierarchical structure, where the nodes at a given level can have an automatically learned weight using techniques similar to the generalized interpolation described in [11].

6. REFERENCES

[1] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," *Proceedings of Eurospeech*, pp. 401–404, 1997.

[2] H.-A. Chang and J. R. Glass, "Hierarchical large-margin Gaussian mixture models for phonetic classification," *IEEE Workshop on ASRU*, pp. 272–277, 2007.

[3] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," *Proceedings of ICASSP*, pp. 313–316, 2007.

[4] E. McDermott, T. J. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007.

[5] A. Park, T. J. Hazen, and J. R. Glass, "Automatic processing of audio lectures for information retrieval: vocabulary selection and language modeling," *Proceedings of ICASSP*, pp. 497–500, 2005.

[6] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp. 453–455, 1994.

[7] H.-A. Chang, "Large-margin Gaissian mixture modeling for automatic speech recognition," M.S. thesis, Massachusetts Institute of Technology, 2008.

[8] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[9] L. Hetherington, "MIT finite-state transducer toolkit for speech and language processing," *Proceedings of ICSLP*, pp. 2609–2612, 2004.

[10] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *Proceesings of ICASSP*, pp. 532–535, 1989.

[11] B. Hsu, "Generalized linear interpolation of language models," *IEEE Workshop on ASRU*, pp. 136–140, 2007.