DATA SAMPLING BASED ENSEMBLE ACOUSTIC MODELLING

Xin Chen and Yunxin Zhao

Department of Computer Science
University of Missouri, Columbia, MO 65211 USA
xinchen@mizzou.edu zhaoy@missouri.edu

ABSTRACT

In this paper, we propose a novel technique of using Cross Validation (CV) data sampling to construct an ensemble of acoustic models for conversational speech recognition. We further propose using Hierarchical Gaussian Mixture Model (HGMM) and repartition training data to increase the ensemble size and diversity. The proposed methods are found to work well together for ensemble acoustic modeling. We also evaluated the quality of the ensemble acoustic models by using the measures of classification margin, average correct score and variance of correct score. We have found that the ensemble of acoustic models increases the margin and the average correct score, and reduces the variance. We compared the performance of our proposed method with a recently reported method of CV Expectation Maximization (CVEM) for single acoustic models. Our experimental results on a telemedicine automatic captioning task showed that the proposed ensemble acoustic modeling has led to significant improvements in word recognition accuracy.

Index Terms— ensemble classifier, acoustic modeling, hierarchical mixture ensemble, data sampling, cross validation

1. INTRODUCTION

Combining multiple speech recognition systems has been established as an effective technique for improving the accuracy performance of automatic speech recognition [1][2][3], where each speech recognition system works independently and the decoding word hypotheses of the multiple systems are combined. Recently, a novel technique has been proposed for generating random-forests based multiple acoustic models and combining their scores for each speech frame [4]. This approach requires only one decoding search for each speech utterance, and therefore it has a significantly lower computation complexity in comparison with the system-level combination.

In machine learning, many methods have been proposed for ensemble classifier design [5], where a noteworthy success is the random forests of decision trees constructed from random samplings on split variables and data [6]. In speech recognition, random sampling on phonetic questions has been used to generate random forests of phonetic decision trees for ensemble acoustic modeling with model-level combination [4], and boosting has been used to generate multiple acoustic models on semi-supervised training data for system-level combination [3]. Along the direction of improving conventional single acoustic models, CV-based data partition has been used in the Expectation Maximization (EM) algorithm to avoid overtraining, where the sufficient statistics of subsets of data were computed by using a model that was estimated from an independent subset of data [7] (this method is referred to as CV-EM).

In this paper, we propose a novel Cross-Validation (CV) data partition based sampling technique to generate an ensemble of acoustic models. We combine these acoustic models in triphone HMM states to compute the scores for each speech frame as in [4]. In addition, we propose two methods to help increase the ensemble size without increasing the overlaps between the sampled training data sets and therefore enhance model diversity. In method one, we use Hierarchical Gaussian Mixture Model (HGMM) to integrate GMDs with different mixture sizes into the ensemble. In method two, for an N-fold CV we repartition the training data set to double the ensemble size to $2N$ while keeping the overlaps between the sampled datasets to be smaller than those in $2N$-fold CV. We also experimentally compared our data sampling method for ensemble acoustic modeling with the method of CVEM, and further, we investigated using CVEM in place of EM in our ensemble acoustic model training. By combining all the methods discussed in this paper, we have obtained the best recognition results on our task of telehealth automatic captioning. With the ensemble size of 60, we have obtained a 3.2% absolute word accuracy gain over our baseline system.

The rest of the paper is organized as the following. In section 2 we introduce the data sampling method for ensemble acoustic model training. In section 3 we describe the two methods for increasing the ensemble size. In section 4 we present the experimental results. In section 5 we analyze the effects of the ensemble acoustic models. In section 6, we conclude our work and discuss possible future extensions.

2. DATA SAMPLING ACOUSTIC MODELING

Context Dependent-Hidden Markov Modeling (CD-HMM) of phoneme units has been proven effective for acoustic modeling. Context-Dependent (CD) phone units are used because acoustic realizations of a phoneme change with its neighboring phonemes. Phonetic Decision Tree (PDT) was proposed in [8] to cluster the allophones in a top-down manner to improve the robustness of CD-HMMs. The root node of a tree contains all the CD-phone data of a phone state. A node is split to two children nodes by using the phonetic context question that produces the largest likelihood gain for the data at the node. Each leaf node of the tree represents a tied state that is shared by similar CD phone states and is modeled by a Gaussian Mixture Density (GMD).

In our proposed method, multiple training data sets are produced through data sampling and each sampled training data set is used to train one set of acoustic models. The CV based data sampling is chosen among other sampling methods because its fixed data usage rate for all the sampled training sets: for an N-fold

CV sampling, a *(N-1)/N* fraction of training data is included in each sampled training set, and each data element is used exactly *N-1* times overall, which help avoid bias in data usage.

The procedure of using *N*-fold CV data sampling to produce an ensemble of acoustic models is shown in Fig.1. In the training data set, training speech data are arranged by the order of their recording time. For an *N*-fold CV, the training set is partitioned into *N* subsets of Group1, 2 … *N*, and the *n*th sampled training set is the training set excluding the Group *n*. The data count for the group partition is based on sentence unit. From each sampled training set, a set of acoustic models of tied triphone HMMs is trained by HTK [9].

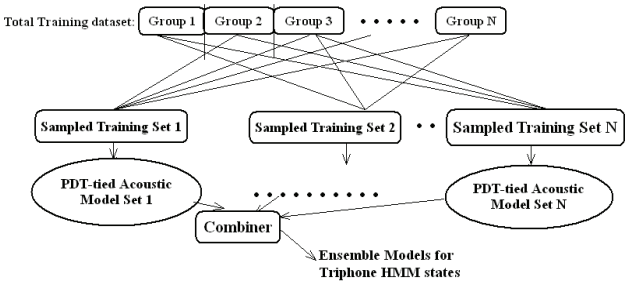

Fig. 1 *N*-fold CV data sampling for ensemble model training

For each triphone HMM state, its tied-state GMDs in the *N* sets of acoustic models are combined to form its ensemble model. As in [4], the triphone HMM states that share the same tied states in every set of acoustic models form a Random-Forests (RF) tied state and have the same ensemble model. In decoding search, the likelihood scores calculated from the GMDs within the same RF-tied state are combined for each speech frame.

The acoustic score combining methods have been discussed in [4][10]. A pilot experiment showed that for data sampling based ensembles, the performance difference between simple average and other weighted average methods is small, and therefore the likelihood scores from the GMDs of each RF-tied state is averaged with uniform weights in the subsequent study.

3. HGMM AND REPARTITION TRAINING DATA

Upon applying an *N*-fold CV data sampling on the training data, we can train *N* sets of acoustic models. To add diversity to the ensemble, we propose the following two methods to increase the ensemble size.

3.1 Hierarchical Gaussian Mixture Model

Mixture size is an important structural parameter in GMD. A model with a small mixture size can be well trained from a small amount of training data but its accuracy may be low. A model with a large mixture size may be accurate but it requires a lot of data to estimate reliably. Therefore a proper mixture size needs to be used for a given training data set. On our telehealth captioning task, the mixture size of 16 worked the best, while increasing the mixture size to 32 led to overfitting.

It has been shown in [4] that ensemble acoustic models are less prone to overfitting and therefore work better with higher mixture sizes. In [11] the authors suggested to use different code book sizes for a multiple HMM based handwriting recognition system to enhance the accuracy performance. Inspired by these works, we proposed to combine models with different mixture sizes into the ensemble of acoustic models so as to integrate the advantages of high accuracy in GMDs of large mixture sizes as well as the robustness in GMDs of small mixture sizes.

3.2 Repartition training data

With an *N*-fold CV, in order to generate more sampled training sets while fixing the *(N-1)/N* usage of training data, one may perform a random shuffling on the training data and from which to obtain new sampled training sets. A disadvantage of the shuffling method is that the sampled sets thus generated will lose the time order in original data, which may be relevant to dialog topics, style, etc. In order to better preserve the time order information in training data, we propose to repartition the training data through shifting the group boundaries as shown in Fig. 2.

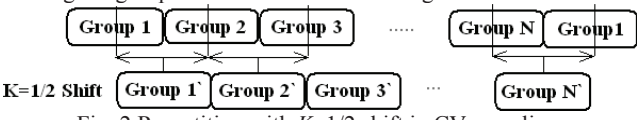

Fig. 2 Repartition with *K*=1/2 shift in CV sampling.

For an *N*-fold CV sampling, with the training data size of *S*, using a shift parameter *K* means to move the start point of the group 1' to *(S/N)*K*.

4. EXPERIMENTAL RESULTS

Experiments were performed on the Telemedicine automatic captioning system developed in the Spoken Language and Information Processing Laboratory (SLIPL) at the university of Missouri-Columbia. The task speech was spontaneous and the vocabulary size was 46k. Please refer to [12] for a detailed description of this task and the system.

4.1 Experiment setup

Speaker dependent acoustic models were trained for 5 health-provider speakers Dr. 1-Dr. 5. Speech features consisted of 39 components including 13 MFCCs and their first and second order time derivatives. Feature analysis was made at a 10 ms frame rate with a 20 ms window size. Gaussian Mixture Density based Hidden Markov Models (GMD-HMM) were used for within-word triphone modeling and the baseline GMDs each had 16 Gaussian components.

4.2 Cross validation data partition based ensemble

We first applied a 10-fold CV data sampling to produce an ensemble of 10 sets of acoustic models, which gave a 2.2% absolute word accuracy gain over the baseline. This performance gain is statistically significant on the telehealth captioning task. For detailed accounts on the significance test on this task, please refer to [4].

Table 1 Word accuracy of the baseline and the 10-fold CV ensemble.

Speakers (Word Counts)	Dr.1 (3248)	Dr.2 (5085)	Dr.3 (3988)	Dr.4 (2759)	Dr.5 (6421)	Average
Baseline	77.43%	81.20%	82.62%	74.12%	78.71%	79.24%
10M_CV	79.09%	83.21%	85.23%	76.48%	81.11%	81.47%

Although the ensemble of 10 acoustic model sets gave a large performance improvement, the individual acoustic model sets had poorer performances than the baseline, indicating that the quality

of individual models was compromised by removing 10% of the training data. For example on Dr.2's data, the averaged word accuracy from the individual acoustic model sets was 80.85%, with a standard deviation of 0.4%. It is clear that the diversity of the acoustic model sets contributed to the overall performance gain, which is an appealing characteristic of ensemble classifiers.

The number of folds in CV is an important parameter for the proposed method. It is straightforward to see that for an *N*-fold CV sampling, each training data set will have *((N-1)/N)* fraction of the total training data, and the overlap between any pair of training data sets is *(1-1/(N-1))*. Therefore, a large *N* implies using more data to train each model set and thus more stable base models, and it also implies larger correlations among the model sets and thus lower diversity in the ensemble. A small *N* would have opposite effects. We performed an experiment to evaluate the effect of the CV parameter *N* on word accuracy, and the results are shown in table 2. The fact that the 5-folds CV ensemble model had the lowest word accuracy may indicate that the diversity was compromised by the instability of the individual model sets.

Table 2 The effect of different fold sizes

	5 fold	10 fold	20 fold
Word Accuracy	80.88%	81.47%	81.36%

4.3 Hierarchical Gaussian Mixture Model

To evaluate the proposed HGMM method, we empirically selected mixture sizes of 16, 24 and 32 as three hierarchical levels. In Table 3 below, the word recognition accuracies are shown for conventional single model sets with the three mixture sizes, as well as for a hierarchical Gaussian mixture model set that is combined from the three sets of acoustic models in the way described in section 2. It is observed that the accuracy performance degraded with the increase of mixture sizes for the single model sets. However, the HGMM gained a 0.46% word accuracy over the best individual model set.

Table 3 The effect of HGMM acoustic model

	16Mix	24Mix	32Mix	3M_HGMM
Word Accuracy	79.24%	77.76%	75.09%	79.70%

We further combined HGMM with the 10-fold CV models to obtain an ensemble that is consisted of 10-CV models with the mixture size of 16, 10-CV models with the mixture size of 24, and 10-CV models with the mixture size of 32, which is referred to as 30M_CV_HIE. In addition, we applied the *K*=1/2 SHIFT on the 30M_CV_HIE model to obtain an ensemble that consists of 20 model sets for each of the three mixture sizes, which is referred to as 60M_CV_HIE_SHIFT. For each of the above model architectures, we also compared using the conventional EM algorithm versus using CVEM for model parameter estimation. For EM, 2 iterations were used and for CVEM, the number of folds was set to 10 and 4 iterations of CVEM were used after 2 iterations of EM. (This implementation choice for CVEM was compared against using CVEM in all iterations in a pilot experiment, which favored the EM+CVEM combination for its smaller computation overhead as well as its good performance.)

Under the same condition of 10-fold CV data sampling, the ensemble of 10-CV acoustic models yielded an average word accuracy of 81.47% in contrast to the 79.72% by CVEM based single model.

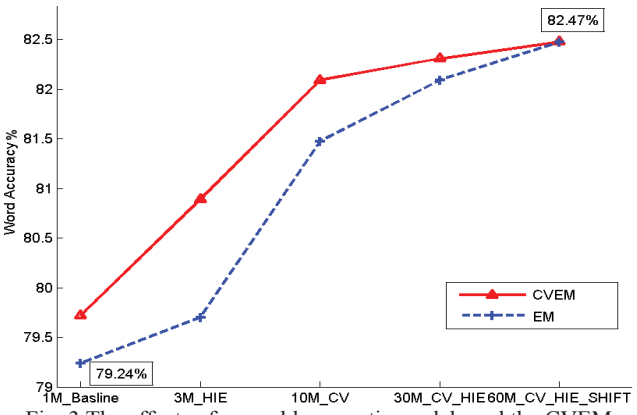

Fig. 3 The effects of ensemble acoustic models and the CVEM incorporated acoustic models

The ensemble of 60 model sets with EM yielded the average word accuracy of 82.47% on the 5 Dr.'s data. It is a 3.2% absolute gain over the baseline. CVEM showed better performance over EM for ensembles with smaller sizes. Its superiority reduced with the increase of the ensemble size, where at the ensemble size of 60, CVEM yielded an average accuracy of 82.48% in comparison with the EM of 82.47%.

5. DISCUSSIONS

5.1 Quality measurement on acoustic models

We used three measures to characterize the quality of the acoustic models. We first define a measure of classification margin:

$$d(x_{i,k}^t) = \log p(x_{i,k}^t \mid \lambda_{C_{i,k}}) - \max_{C_j \in \Omega, C_j \notin Phone(C_j)} \log p(x_{i,k}^t \mid \lambda_{C_{j,k}})$$

where $x_{i,k}^t$ represents a speech frame at time t of the class $C_{i,k}$ corresponding to triphone i and state k, $\lambda_{C_{j,k}}$ denotes the HMM for the class of triphone j and state k, $Phone(C_j)$ defines the subset of triphone states that share the same center phone, and Ω defines all triphones. Since our focus is on the quality of the acoustic models, the prior $P(C_i)$'s are assumed uniform.

While the margin measure characterizes model quality for class discrimination, we used the averaged correct score per frame to measure the model-data fit, and we also used the averaged standard deviation of the correct frame scores for each triphone class to measure the score variations. The measures were computed separately on the train and test sets of the speaker Dr.2, and the results are given in Table 4. Note that here only the model 1M_16_CVEM used CVEM in parameter estimation, and the 5 other models all used EM. It is observed that our proposed ensemble acoustic models improved the model quality in all of the three aspects, i.e. increased classification margin, increased average correct score, and decreased variance of the correct scores.

Table 4 Acoustic model quality measured on Dr.2's data sets

Dr.2's data	Margin		Average Score		Standard Deviation	
	train	test	train	test	train	Test
1M_32mix	-2.12	-5.34	-102.44	-105.20	21.02	21.10
1M_16_EM	-0.87	-4.59	-102.65	-103.64	20.97	20.84
1M_16_CVEM	-0.86	-4.58	-102.25	-103.24	21.16	20.69
3M_HIE	-1.77	-4.09	-100.90	-102.74	20.50	20.51
10M_CV	1.17	-3.28	-98.71	-100.40	20.06	19.99
30M_CV_HIE	1.14	-2.60	-96.05	-99.35	19.56	19.66

In Fig.4 we show the correct and the competing acoustic scores for one test sentence. It is clear that from the frame 220 to the frame 280 the baseline has large negative margins while the margins are small for the 10CV model. This difference explains the fact that the 10CV model ensemble produced the correct hypothesis "prevent", while the baseline model gave the error hypothesis "prun vent".

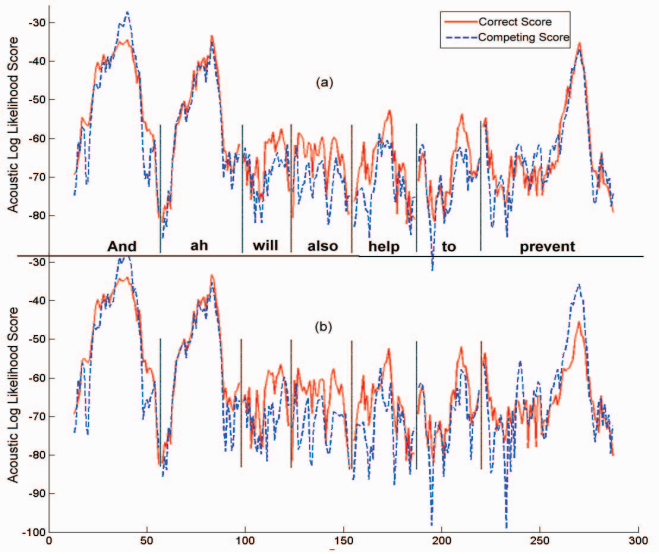

Fig. 4 The correct and the competing scores from one test sentence. (a) 10-fold CV model ensemble (b) baseline model

5.2 Comparison between HGMM and data repartition

Since the methods of repartitioning training set (shift) and HGMM both help increase the ensemble size, we conducted an experiment to compare their effects on accuracy performance with the ensemble size fixed to 20 acoustic model sets. The results are shown in table 5.

Table 5 Experiment on HGMM and data repartition

	HGMM 16mix_24mix	HGMM 16mix_32mix	10CV and 10 K=1/2 Shift CV
20M_EM	81.98%	82.05%	81.46%

It is observed that the ensemble with HGMM is better than the shift of data repartition alone. However, as shown in Fig.3, when applying the shift on CV+HGMM we obtained a large increase in word accuracy, indicating that the two methods have complementary merits and thus are combinable.

5.3 Sampling methods and sampling units

Two issues in the data sampling methods are worth mentioning. The first is on the sampling unit. We performed data partition based on counts of sentences and phonemes, and found insignificant differences in the resulting accuracy performance. We also performed experiment on random sampling without replacement with a 90% data usage. With the same ensemble size, the CV based data sampling worked slightly better than random sampling.

6. CONCLUSION AND FUTURE WORK

Ensemble acoustic modeling is a promising new direction to improve the accuracy performance of ASR. In the current work we have proposed several ensemble acoustic modeling methods, including data sampling, hierarchical mixture, and repartition the training data. By combining the three methods, we achieved a 3.2% absolute gain in word accuracy over our baseline, which is larger in comparison with the feature sampling based RF-PDT ensemble on the same task [4].

Within the framework of our multiple acoustic models based speech recognition, there are a number of potential future extensions. One issue worthy of investigation is optimizing data sampling or partition for different types of ASR tasks. Another extension is to combine data sampling with discriminative training such as MMIE, MCE, and MPE which are successful in acoustic modeling but may overfit to the training data, where through the combination, the diversity of data sampling ensemble may well compensate for the overfitting in the individual models. A further extension is to better utilize the state-of-art multiple-core computer architecture to carry out parallel computation of the acoustic likelihood scores from the models in the ensemble to speed up decoding search (an alternative multiple-pass approach is to use the baseline model to generate a word lattice and to apply the ensemble model for rescoring). A yet further extension is to incorporate automatic determination of language model scale parameters into the speech recognition system, since the increase in average acoustic score in the ensemble acoustic model indicates improved quality of acoustic model, and thus the weights on acoustic model and language model scores should be rebalanced.

7. ACKOWNLEDGEMENT

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC004340-05A2. The authors would like to thank Dr. Jian Xue for his help with the captioning systems, and to thank Dr. Takahiro Shinozaki for his help in CVEM implementation.

8. REFERENCE

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", Proc. IEEE ASRU Workshop, pp. 347-352, 1997.

[2] O. Siohan, B. Ramabhadran and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," Proc. ICASSP, pp. I-197 - I-200, 2005.

[3] R. Zhang, et al., "Investigations on ensemble based semi-supervised acoustic model training," Proc. EuroSpeech, pp. 1677-1680, 2005.

[4] J. Xue, Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," IEEE Trans. SAP, vol.16, iss. 3, pp. 519-528, 2008.

[5] T. G. Dietterich, "Ensemble methods in machine learning," Proc. of the First International Workshop on Multiple Classifier Systems, pp. 1-15, 2000.

[6] L. Breiman "Random forests," Mach. Learn., vol. 45, pp.5-32, 2001.

[7] T. Shinozaki, M. Ostendorf, "Cross-validation EM training for robust parameter estimation," Proc. ICASSP, vol. IV, pp. 437–440, Hawaii, 2007.

[8] S. J. Young, J. J. Odell and P.C. Woodland, "Tree-based state tying for high accuracy modeling," Proc. ARPA Human Lang. Tech. Workshop, pp. 307-312, 1994.

[9] HTK Toolkit, U.K. http://htk.eng.cam.ac.

[10] X. Chen, "Ensemble methods in large vocabulary continuous speech recognition," Master thesis, University of Missouri, 2008.

[11] A. Hung-Ren Ko, et al., "A new HMM based ensemble generation method for numerical recognition," MCS workshop, pp. 52-61, 2007.

[12] Y. Zhao, et al., "An automatic captioning system for telemedicine," Proc. ICASSP, pp. I-957 – I-960, 2006.