CLUSTER-BASED ADAPTATION USING DENSITY FOREST FOR HMM PHONE RECOGNITION

Mohamed Abou-Zleikha[1], *Zheng-Hua Tan*[1], *Mads Græsbøll Christensen*[2], *Søren Holdt Jensen*[1]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Audio Analysis Lab, ad:mt, Aalborg University, Denmark
{moa,zt,shj}@es.aau.dk, mgc@create.aau.dk

ABSTRACT

The dissimilarity between the training and test data in speech recognition systems is known to have a considerable effect on the recognition accuracy. To solve this problem, we use density forest to cluster the data and use maximum a posteriori (MAP) method to build a cluster-based adapted Gaussian mixture models (GMMs) in HMM speech recognition. Specifically, a set of bagged versions of the training data for each state in the HMM is generated, and each of these versions is used to generate one GMM and one tree in the density forest. Thereafter, an acoustic model forest is built by replacing the data of each leaf (cluster) in each tree with the corresponding GMM adapted by the leaf data using the MAP method. The results show that the proposed approach achieves 3.8% (absolute) lower phone error rate compared with the standard HMM/GMM and 0.8% (absolute) lower PER compared with bagged HMM/GMM.

Index Terms— ensemble acoustic modeling, density forest, cluster-based adaptation, HMM speech recognition

1. INTRODUCTION

Hidden Markov model is the mainstream technology for speech recognition in both academy and industry communities. Much research has been conducted for this modelling method to provide more efficient training techniques [1, 2], different types of models (e.g. subspace HMM [3]), and different types of adaptation methods (e.g. supervised and unsupervised adaptation [4, 5]).

The dissimilarity between the training and test data is one of the major factors that affect on the recognition accuracy. Supervised speech recognition adaptation is considered limited and costly in several applications due to its requirement of labeled data. However, an unsupervised adaptation represents an alternative, and would add a considerable improvement to the speech recognition performance.

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

Ensemble machine learning techniques such as the random forest have become a very attractive research direction. At the same time, combining multiple recognition systems is widely used to improve the recognition performance [6, 7, 8, 9, 10, 11, 12], since, in principle, combining the decision of multiple models can give a better performance than only using one. Several attempts have been conducted to use ensemble approaches for speech recognition [7, 9, 10, 11]. These attempts can be classified into two directions. The first one focuses on modelling the phonetic decision trees in the context of context-dependent speech recognition. This is mainly done by generating a random forest of phonetic decision trees and using these trees to generate multiple acoustic models from the training data, and at the recognition phase the output of each model is calculated and the results are combined [9]. The second direction focuses on generating ensemble-based acoustic models. In [10], the authors proposed a cross-validation data sampling approach to generate multiple training data sets through data sampling, and then use each of these sets to train one set of acoustic models. In [11], the authors proposed to use bootstrap and restructuring for hidden Markov acoustic modeling with sparse training data for low resourced languages . [13] proposed to use a bagging approach to train several GMMs per state and then take the average of these GMMs output probabilities. In spite of the promising results the ensemble-based acoustic models achieved, they are still restricted to using the bagging principle only.

The main contribution of this paper is to propose a new cluster-based adaptation technique for speech recognition by investigating the applicability of the density forest as an ensemble data clustering technique combined with maximum a posteriori (MAP) method to build a cluster-based adapted GMM in HMM speech recognition. The motivation of using this approach is to benefit from the robustness of the ensemble machine learning techniques, not only in generating multiple models but also in creating a cluster-based unsupervised adaptation, which can improve the performance of the system.

Using a bagged version of the training data for a state in a

phone HMM, the proposed approach aims to divide this data into smaller groups using the density forest approach, and then through MAP adaptation, update a GMM model that is trained using the bagged version. These cluster-based adapted models are used to replace the leaf data, generating an acoustic model forest. When a new speech frame is presented, the corresponding likelihood from each tree in the forest is calculated, and the final likelihood is the average of all trees likelihoods.

This paper is organised as follows: in Section 2, the density forest and the functions used for information gain calculation are presented. In Section 3 the maximum a posteriori method is briefly explained. The proposed approach is presented in Section 4. In Section 5, the experimental results on the TIMIT speech database are provided, and section 6 presents the conclusion and future work.

2. DENSITY FOREST

Density forest is an ensemble tree-based clustering approach [14]. It is considered as an unsupervised version of the random forest classification technique. The principle is to grow an ensemble of trees on a random selection of samples in a training set. While constructing the trees, at each tree node, randomly selected features is investigated as a potential predictor that decides the split of the data in the tree. These features split the data into two parts; each of these parts is modelled using a density model (e.g. GMM). The splitting fitness is calculated as the information gain generated by the splitting. The tree leaves contain subsets of the tree training data. These data are used to generate prediction models such as GMMs or to modify a model generated using the whole tree training data.

Formally, a density forest R is a set of decision trees

$$R = t_1, t_2, ..., t_{ntree} \tag{1}$$

where t_i is the i^{th} individual tree and $ntree$ is the number of trees.

Given a bagged version of the training data, each tree in the forest is trained independently. Given a set of features $F = f_1...f_n$ (i.e. the acoustic features of each frame), the j_{th} node is split by using the feature that maximises the information gain:

$$f_j = \arg\max_{f \in F_j} I(X_j, f) \tag{2}$$

where F_j is a randomly selected feature subset of F at node j, X_j is the data at node j and $I(X_j, f)$ is the information gain function.

In the density forest, each tree grows until one of the following conditions is achieved:

1. The maximum depth is reached.
2. The number of samples in the node is smaller than a threshold.
3. The information gain is smaller than a threshold.

Several splitting criteria exist. In this paper, three of them are examined for phone recognition task: unsupervised entropy, Kullback-Leibler divergence function and normalised L2 distance function as described in next subsections.

2.1. Unsupervised Entropy

Suppose we have a collection of data points X_j, one which GMM λ_j is trained, we split X_j into two parts X_j^L and X_j^R and then use them to train GMMs λ_j^L and λ_j^R, respectively. Using a feature f and a threshold, the information gain of this splitting is defined as [14]:

$$I(X_j, f) = \log(|\Lambda(\lambda_j)|) - \sum_{i \in L,R} \frac{|X_j^i|}{|X_j|} \log(|\Lambda(\lambda_j^i)|) \tag{3}$$

where $\Lambda(\lambda_j)$ is associated covariance matrix with λ_j and $|\cdot|$ indicates a determinant for a matrix or the number of data points (more information about this function can be found in [14]).

2.2. Kullback-Leibler Divergence

Kullback-Leibler divergence is a mathematical measurement of the difference between two probability distributions. It is calculated as:

$$d_{KL}(p_1, p_2) = \int p_1(x) \log(\frac{p_1(x)}{p_2(x)}) dx \tag{4}$$

where p_1 and p_2 are two probability distributions (i.e. GMMs) estimated from splitting X_j^L and X_j^R, respectively. This function is non-symmetric. In this work, a symmetrised version is used, which is defined as:

$$d_{sKL}(p_1, p_2) = d_{KL}(p_1, p_2) + d_{KL}(p_2, p_1) \tag{5}$$

For Gaussian mixtures, a closed form expression for $d_{KL}(p_1, p_2)$ only exists for the number of mixtures $M = 1$. For $M > 1$, $d_{KL}(p_1, p_2)$ is estimated using stochastic integration or an approximation as described in [15].

2.3. Normalized L2 Distance

The normalised L2 distance function is defined as [16]:

$$d_{nL2} = \int (p_1^{'}(x) - p_2^{'}(x))^2 dx \tag{6}$$

where p_1 and p_2 are two probability distributions (i.e. GMMs) estimated from X_j^L and X_j^R, respectively, and $p_i^{'}(x) = p_i(x) / \sqrt{\int p_i(x)^2 dx}$ is a scaled form of $p_i(x)$ to unit L2-norm.

2.4. The Maximum a Posteriori Method

The maximum a posteriori estimation method is a powerful approach for updating GMMs [17]. This method was initially proposed for speaker adaptation in speech recognition. Adaptation can be applied to all or to a number of the GMM parameters. In this paper, the update is only applied to the means of GMMs.

Given a set of samples called the adaptation data $X = x_1, ..., x_T$, and the GMM model $\lambda = (w_k, \mu_k, \Sigma_k)_{k=1}^{M}$, the adapted mean vector $\acute{\mu}$ is calculated as the weighted sum of the adaptation data and the GMM mean as:

$$\acute{\mu}_k = a_k \tilde{x}_k + (1 - a_k)\mu_k \tag{7}$$

where

$$a_k = \frac{n_k}{n_k + r} \tag{8}$$

$$\tilde{x}_k = \frac{1}{n_k} \sum_{t=1}^{T} P(k|x_t) x_t \tag{9}$$

$$n_k = \sum_{t=1}^{T} P(k|x_t) \tag{10}$$

$$P(k|x_t) = \frac{w_k p_k(x_t))}{\sum_{j=1}^{M} w_j p_j(x_t)} \tag{11}$$

where M is the number of mixtures, $p_i(x)$ denotes Gaussian probability and r is a constant controlling the influence of the prior probability.

3. THE PROPOSED APPROACH

In the proposed approach, multiple training data sets are generated by bagging the training data for a state in a phone HMM. Each of these sets is used to generate a GMM model and a density tree, and then the data at the tree leaves are used to update the GMM model to generate an acoustic model tree.

In the following, the procedure of training the forest is described.

- Extract the training data for a state in a phone HMM using HTK [18].
- For each tree in the forest
  - Generate a bagged version of the phone state training data X.
  - Density Tree Builder: a density tree is generated using the bagged training data X.
  - GMM Builder: a GMM is built using X.
  - For each leaf in the tree, the data in that leaf is used to update the corresponding generated GMM using the MAP method.

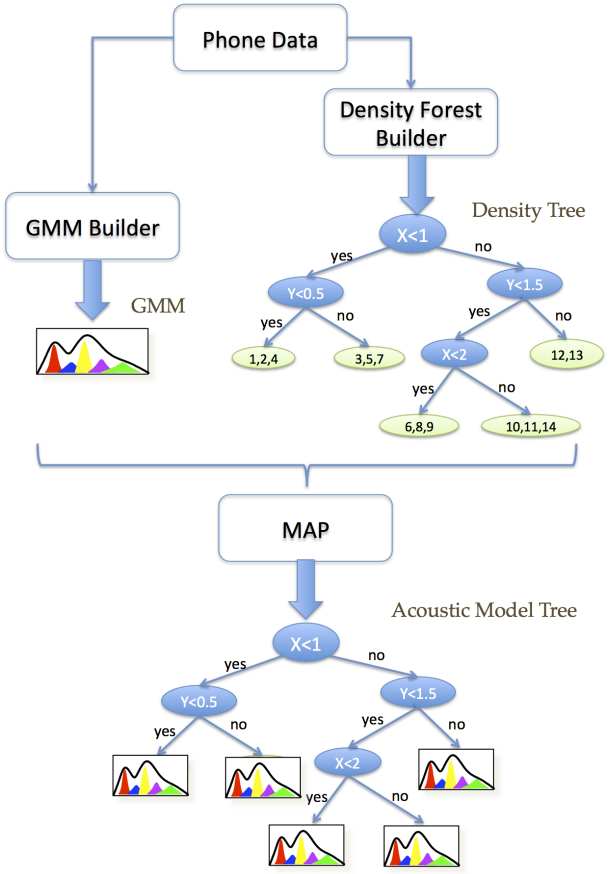

Fig. 1. The training procedure of a tree in the density forest.

Figure 1 illustrates the procedure of building one tree. As a result of the procedure, a set of trees is generated. The leaves of each tree are GMM models. We call this forest an acoustic model forest. At the recognition phase, for each speech frame, each tree in the forest is traversed to retrieve the adapted GMM (one GMM per tree).Then the likelihood for each GMM is calculated, and the final frame likelihood is the average of the likelihoods for all these GMMs (trees).

4. EXPERIMENTS AND EVALUATION

To evaluate the proposed approach, phoneme recognition experiments have been conducted on the TIMIT corpus [19]. The standard training set has been used after excluding all SA records (i.e. identical sentences for all speakers) as they might bias certain phoneme contexts and result in artificially high recognition scores [20]. Results are reported for the core test set. Speech feature are 12 Mel Frequency Cepstral Coeficients (MFCC) and the logarithm of the energy together with their first and second temporal derivatives. In training phase, the 61 phones are mapped to 48 phones and in testing phase, the 61 phones are mapped to 39 phones [20].

HTK is used to retrieve the training data for a state in

a phone HMM and the transition probabilities using a standard 3-state context independent monophone HMM recogniser with 32 GMM components. The proposed approach has been compared with two systems, the standard 3-state context independent monophone HMM recognisers with 32 GMM components, and a bagging-based ensemble system with 25 generated models where 25 training sets are produced from the training data for a state in a phone HMM and are used to build a set of GMMs (a context independent realisation of system described in [13]).

For the proposed approach, the number of trees in the forest is 25, the maximum depth of forest trees is fixed to 10, and the number of randomly examined features per node is 13. The bagging is done using random sampling with replacement approach.

The phoneme recognition experiments were conducted. Table 1 presents the phone error rates of the proposed approach using the three information gain functions and of the baseline systems. The terms DF-UE, DF-KL and DF-L2 represent the method proposed using the unsupervised entropy, Kullback-Leibler divergence and normalized L2 distance as splitting criteria.

Table 1. Phone error rates of the proposed approach using each examined information gain function and of the baseline systems.

Method	Phone Error Rate (%)	Correctness %
Standard HMM	32.4	73.4
Bagged HMM	29.4	74.3
DF-UE	29.0	74.4
DF-KL	**28.6**	74.4
DF-L2	29.1	74.0

Table 1 shows that the proposed approach has outperformed both standard HMM and the bagged HMM using the three splitting criteria. The lowest phone error rate has been achieved using Kullback-Leibler divergence function as an information gain function, with 28.6% error rate, comparing with 32.4% error rate for standard HMM and 29.4% for the bagged HMM. The reason for the difference in performance between the proposed approach and the bagged HMM one is that in addition to the bagging, a clustering is performance and a model adaptation is applied using the data of each cluster. On testing phase, the model is selected according to the distance similarity between the testing data and the model data using the density forest.

To check which error aspects have been reduced, a comparison between the insertion, deletion and substitution have been performed between the best result obtained from the proposed approach and bagged HMM as shown in Table 2. The results show that the main advantage is achieved on insertion and substitution with 0.9% less. However, the deletion error has been increased by 0.8%.

Table 2. Insertion, Deletion and Substitution percentage in the proposed proposed approach and the bagged HMM

Method	Insertion %	Deletion %	Substitution %
Bagged HMM	3.9	7.6	18.1
DF-KL	3.0	8.4	17.2

To examine the effect of the proposed approach on each phone, a single phone error rate (SER) is calculated for each of the 48 phones (without mapped to 39 phones). Table 3 shows the SER values using the DF-KL approach and the Bagged HMM approach. The results show that most of the phones (marked in bold) get benefits from the clustering proposed.

Table 3. Single phone error rates of the proposed approach using KL information gain function and the bagged HMM.

Phone	Bagged HMM	DF-KL	Phone	Bagged HMM	DF-KL
aa	0.78	0.6	**iy**	0.46	0.3
ae	0.74	0.5	**jh**	0.6	0.31
ah	0.93	0.69	**k**	0.44	0.28
ao	0.71	0.44	**l**	0.59	0.45
aw	0.97	0.9	**m**	0.44	0.32
ax	0.91	0.7	**n**	0.46	0.37
ay	0.53	0.42	**ng**	0.73	0.69
b	0.55	0.42	**ow**	0.57	0.56
ch	0.73	0.73	oy	1	1.19
cl	0.38	0.29	**p**	0.51	0.43
d	0.71	0.6	**r**	0.54	0.47
dh	0.68	0.55	**s**	0.31	0.22
dx	0.57	0.44	**sh**	0.49	0.25
eh	0.86	0.67	sil	0.07	0.08
el	0.95	0.82	**t**	0.57	0.37
en	1.14	1.1	**th**	1.11	1.08
epi	1.06	1	uh	1.07	1.24
er	0.48	0.43	**uw**	0.72	0.7
ey	0.52	0.39	**v**	0.59	0.42
f	0.44	0.27	**vcl**	0.56	0.48
g	0.77	0.79	**w**	0.47	0.39
hh	0.62	0.44	**y**	0.88	0.78
ih	0.79	0.67	**z**	0.62	0.43
ix	0.62	0.56	zh	1.1	1.3

The complexity of the proposed approach can be investigated by studying the relationship between recognition time and the number of trees. These two factors were found to be correlated where the complexity is defined as $O(N)$ where N is the number of trees in the acoustic model forest, which makes the proposed approach has the same complexity of the bagged HMM.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a tree-based ensemble method for a cluster-based adaptation in HMM/GMM phone recognition has been proposed. We investigated the applicability of the density forest as an ensemble data clustering technique combined with MAP method to build a cluster-based adapted GMM in HMM speech recognition. As a result, an acoustic model forest is generated, where each leaf in each tree in that forest represents one GMM. At the recognition phase, for each frame, the likelihood from each tree is calculated, and the final frame likelihood at that state is the average of the likelihoods of all trees. Three gain functions for building model forest has been examined, unsupervised entropy function, the Kullback-Leibler divergence function and the normalised L2 distance function.

Phone recognition error rates on TIMIT corpus show that the proposed approach has outperformed both the standard HMM and the bagged HMM using the three splitting criteria. The best performance is achieved using Kullback-Leibler divergence as a gain information function for data splitting in the density forest.

Future work includes investigating the effect of number of trees on the performance, the usage of different GMM adaptation technique instead of MAP, and finding different strategies for combining the likelihood than averaging.

6. REFERENCES

[1] F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden markov models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, 2007, pp. IV–313.

[2] J. Keshet, C. Cheng, M. Stoehr, A. McAllester, and sK. Saul, "Direct error rate minimization of hidden markov models.," in *INTERSPEECH*, 2011.

[3] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N.K. Goel, M. Karafiát, A. Rastrow, et al., "Subspace gaussian mixture models for speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4330–4333.

[4] C. Lee and J. Gauvain, "Speaker adaptation based on map estimation of hmm parameters," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1993, pp. 558–561.

[5] L. Machlica, Z. Zajíc, and A. Pražák, "Methods of unsupervised adaptation in online speech recognition," *SPECOM'2009 Proceedings*, 2009.

[6] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Workshop on Automatic Speech Recognition and Understandings*, 1997, pp. 347–354.

[7] P. Xu and F. Jelinek, "Random forests in language modeling," in *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[8] H. Xu, Z.H. Tan, P. Dalsgaard, and B. Lindberg, "Robust speech recognition based on noise and snr classification-a multiple-model framework," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 977–980.

[9] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *Transactions on Audio, Speech, and Language Processing*, pp. 519–528, 2008.

[10] X. Chen and Y. Zhao, "Data sampling based ensemble acoustic modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2009, pp. 3805–3808.

[11] Xiaodong Cui, Jian Xue, Xin Chen, Peder A Olsen, Pierre L Dognin, Upendra V Chaudhari, John R Hershey, and Bowen Zhou, "Hidden markov acoustic modeling with bootstrap and restructuring for low-resourced languages," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2252–2264, 2012.

[12] X. Chen and Y. Zhao, "Building acoustic model ensembles by data sampling with enhanced trainings and features," *Transactions on Audio, Speech, and Language Processing*, 2013.

[13] C. Dimitrakakis and S. Bengio, "Phoneme and sentence-level ensembles for speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2011.

[14] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning," *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, p. 12, 2011.

[15] E. Pampalk, "Speeding up music similarity," *2nd Annual Music Information Retrieval eXchange, London*, 2005.

[16] J. H. Jensen, D. Ellis, M. G. Christensen, and S. H. Jensen, "Evaluation distance measures between gaussian mixture models of mfccs," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007, pp. 107–108.

[17] Chin-Hui Lee, C-H Lin, and Biing-Hwang Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 806–814, 1991.

[18] "HTK toolkits, Cambridge, UK," http://htk.eng.cam.ac.

[19] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," *Speech Input/Output Assessment and Speech Databases*, 1989.

[20] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, pp. 1641–1648, 1989.