2010 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2010

March 14-19, 2010 • Dallas, Texas, U.S.A.

IEEE Catalog Number: CFP10ICA-CDR
ISBN: 978-1-4244-4296-6 ISSN: 1520-6149

GMM-HMM ACOUSTIC MODEL TRAINING BY A TWO LEVEL PROCEDURE WITH GAUSSIAN COMPONENTS DETERMINED BY AUTOMATIC MODEL SELECTION

Dan Su[1], *Xihong Wu*[1], *and Lei Xu*[1,2]

[1]Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education), Peking University
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong

ABSTRACT

This paper investigates the Bayesian Ying-Yang (BYY) learning for speech recognition via Gaussian mixture models (GMMs) based Hidden Markov models (HMMs). A two level procedure is proposed with the hidden Markov level trained still under the maximum likelihood principle by the Baum-Welch algorithm but with the GMMs level trained under the BYY best harmony. We proposed a new batch way EM-like Ying-Yang alternation algorithm and used it as a plug-in block to the Baum-Welch algorithm. The advantage is that number of GMM components can be automatically determined during this BYY harmony learning and that the resulted model parameters become less affected than EM-ML training by the problem of overfitting and singular solution. In comparison with the standard EM-ML training and classical model selection criterions, including BIC and AIC, speech recognition experiments in a large vocabulary task on the Hub4 broadcast news database shown that the proposed algorithm provides an improved performance and also good convergence.

Index Terms— speech recognition, model selection, Bayesian Ying-Yang learning, GMMs, HMMs

1. INTRODUCTION

The acoustic model in modern ASR systems has a very complicated structure: hidden markov level is composed of a set of clustered states and each state's output distribution is represented by a multivariate GMM. Improvements on performance can be expected if optimal model size can be determined and thus more precise model parameters can be estimated. Some work has been done to prune mixture components using classical model selection techniques like the Akaike information criterion (AIC)[1], Bayes information criterion (BIC) [2], and minimum description length (MDL) [3]. However, there are two problems when apply them into ASR systems, first, with a two stage implementation, they typically need to enumerate all possible candidate number of components; second, the estimation performance deteriorates when data dimension is high and the number of parameters is large. The two problems have made them impractical for speech recognition system trained on large databases. Recently, the application of Variational Bayesian (VB) approach for speech recognition systems has also been investigated [4, 5] and achieve improvement on recognition accuracy, but the performance seems sensitive to the choice of priors.

BYY learning [8, 9] is a relatively new learning technique that allows model parameters and model scale be learned simultaneously and automatically.The problems for BIC and AIC can be avoided in the BYY approach, in which extra components are automatically pruned during training. To fully adopt BYY learning to generate more precise acoustic model parameters, the whole framework can be involved, that is, both model size at hidden Markov level and Gaussian mixture level can be determined. Instead of applying BYY learning for both levels at one time, as a first step, also for convenient comparison with classical model selection criterions, model selection by BYY is primarily applied to the GMM level to determine the number of Gaussian components in this work.

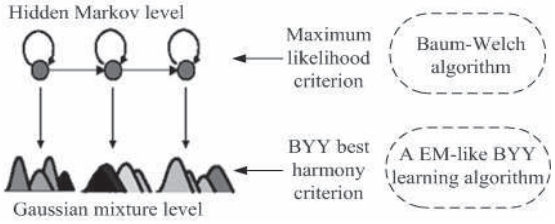


Fig. 1. The training framework of a two-level procedure .

We propose a two level training procedure in which the hidden Markov level is still trained under maximum likelihood principle with Baum-Welch (BW) algorithm, while the GMM level is trained under the BYY best harmony principle. The framework is illustrated as in Figure 1. Instead of directly adopting the existing adaptive BYY learning algorithm for GMMs, we proposed a new EM-like BYY learning algorithm in help of an auxiliary function based regularization. The proposed algorithm is compatible with BW algorithm in that parameters are updated in batch mode. We have mathematically proved that by appropriately selecting a smoothing parameter the algorithm can ensure convergence of harmony functional. Experiments on large vocabulary Hub4 Mandarin speech corpus has shown its effectiveness and good convergence.

The rest of this paper is organized as follows. Section 2 describes our proposed training method for acoustic models. Experiments are presented in section 3, and conclusion is made in section 4.

2. PROPOSED TRAINING ALGORITHM FOR GMM-HMM ACOUSTIC MODEL

2.1. BYY learning formulation for GMMs

Firstly proposed in [8] and systematically developed over a

decade [9], BYY harmony learning theory is a general statistical learning framework that can handle both parameter learning and model selection under a best harmony principle. A salient advantage of BYY learning is that model selection can be done simultaneously with parameter learning.

The general form of harmony measure and its specific form for GMMs are restated in Eq (1) and (2). The details are referred to [10]) and especially a recent systematic tutorial in [11]. In Eq. (2) the smoothing item h is ignored:

$$H(p||q,\theta) = H_f(p||q,\theta) - Z(h) \tag{1}$$

$$H_f(p||q,\theta) = \sum_{t=1}^{T}\sum_{j=1}^{k} p(j|x_t,\theta)\ln q(x_t,j|\theta) \tag{2}$$

in which $q(x_t,j|\theta) = \alpha_j \mathcal{N}(x_t|\mu_j,\Sigma_j)$ and $p(j|x_t,\theta)$ can be set with Bayesian structure, i.e. $p(j|x_t,\theta) = \frac{q(x_t,j|\theta)}{\sum_k q(x_t,k|\theta)}$.

To maximize $H_f(p||q,\theta)$, existing adaptive algorithm is implemented via a Ying-Yang iteration procedure, in which at Yang-step, $p(j|x_t,\theta)$ and δ_{tj} in Eq. (3) are calculated according to above equations, then, at Ying-step, model parameters are updated with gradient based approach.

$$\nabla_{\theta_j} H_f(X,\theta,k) = \sum_t (1+\delta_{tj}) p(j|x_t,\theta)\frac{\partial \ln q(x_t,j|\theta)}{\partial \theta_j} \tag{3}$$

where $\delta_{tj} = \ln p(j|x_t,\theta) - \sum_{l=1}^{k} p(l|x_t,\theta)\ln p(l|x_t,\theta)$.

To be applied for GMM level in acoustic model, we propose an EM-like BYY learning algorithm to embed into the Baum-Welch training framework for hidden Markov state level since the BW algorithm is typically in batch way and also free from determination of any learning rate parameter.

2.2. Proposed EM-like BYY Learning Algorithm

Parameter estimation for GMMs under ML criterion is considered a well solved problem based on the EM technique. This is an iterative procedure in which each iteration is a two-step process, and which is guaranteed to converge to a local optimum. The first step involves accumulating statistics which depend on the current estimated distribution of a hidden variable, the second step maximizes a so called "auxiliary function", if its value is increased, the object function is bound to increase too. Moreover the auxiliary function should be easier to be directly maximized than the object function.

In the object function of best harmony criterion, the parameters exist in both Ying part and Yang part and need to be estimated simultaneously. It is not straightforward to derive an auxiliary function which is guaranteed to increase the objective function to perform an EM-like updating. To solve this problem, we adopt the method used in [6, 7]. In this method, instead of finding a strong sense auxiliary function, a weak sense auxiliary function and a smoothing function are constructed.

A weak-sense auxiliary function $G_{weak}(\theta,\tilde{\theta})$ has the same gradient with the object function around the point $\theta = \tilde{\theta}$. For the BYY harmony functional, a weak-sense function can be naturally written according to Eq. (3):

$$G_{weak}(\theta,\tilde{\theta}) = \sum_t\sum_j (1+\tilde{\delta}_{tj}) p(j|x_t;\tilde{\theta})\ln q(x_t,j;\theta)$$

$$\text{with } \tilde{\delta}_{tj} = \ln p(j|x_t,\tilde{\theta}) - \sum_{l=1}^{k} p(l|x_t,\tilde{\theta})\ln p(l|x_t,\tilde{\theta}) \tag{4}$$

A smoothing function has its maximum at current point $\theta = \tilde{\theta}$. It can be added to a weak-sense auxiliary function to improve convergence. For the BYY harmony functional, a possible form of smoothing function can be as following, which satisfies $\left.\frac{\partial G_{sm}(\theta,\tilde{\theta})}{\partial\theta}\right|_{\theta=\tilde{\theta}} = 0$:

$$G_{sm}(\theta,\tilde{\theta}) = -0.5\sum_j D_j\{\ln|\Sigma_j| + Tr[(\tilde{\Sigma}_j + \tilde{S}_j)\Sigma_j^{-1}]\},$$

where $\tilde{S}_j = (\mu_j - \tilde{\mu}_j)(\mu_j - \tilde{\mu}_j)^T$ and D_j is a constant that controls the amount of parameters to be smoothed.

The two functions together, target the same property as a strong-sense auxiliary function, thus the following auxiliary function can be constructed:

$$F(\theta,\tilde{\theta}) = G_{weak}(\theta,\tilde{\theta}) + G_{sm}(\theta,\tilde{\theta}) \tag{5}$$

Maximization of Eq. (6) leads to the following EM-like Ying-Yang iteration:

- Yang-step: calculate $p(j|x_t,\theta)$ and δ_{tj}
- Ying-step: updating α_j, μ_j and Σ_j as follows

$$\alpha_j = \sum_t \xi_{tj} / \sum_l\sum_t \xi_{tl} \tag{6}$$

$$\mu_j = \frac{\sum_t \xi_{tj}x_t + D_j\tilde{\mu}_j}{\sum_t \xi_{tj} + D_j} \tag{7}$$

$$\Sigma_j = \frac{\sum_t \xi_{tj}[(x_t-\mu_j)(x_t-\mu_j)^T] + D_j\tilde{\Sigma}_j}{\sum_t \xi_{tj} + D_j}, \tag{8}$$

in which $\xi_{tj} = (1+\delta_{tj})p(j|x_t,\theta)$.

For mixture weight parameters, a specific smoothing function can also be constructed, then another smoothing constant will be introduced and by appropriately setting it the positive of α_j can be ensured. However, preliminary experiments show that there is no obvious effect on performance by adding smoothing function for the mixture weight parameters. Therefore, in this work, we simply update mixture weight parameters with Eq. (6), when α_j falls negative, the corresponding component is simply discarded.

It can be proved that by selecting appropriate D_j, the above updating rule ensure positive projection on the gradient both for μ_j and Σ_j (when Σ_j is diagonal), see following:

$$vec[\mu_j - \tilde{\mu}_j]^T \cdot vec[\nabla_{\mu_j} H] = \frac{1}{N(\sum_t \xi_{tj} + D_j)}\cdot$$
$$tr[(\sum_t \xi_{tj}x_t - \sum_t \xi_{tj}\tilde{\mu}_j)^T\Sigma_j^{-1}(\sum_t \xi_{tj}x_t - \sum_t \xi_{tj}\tilde{\mu}_j)]$$

$$vec[\Sigma_j - \tilde{\Sigma}_j]^T \cdot vec[\nabla_{\Sigma_j} H] = \frac{0.5}{N(\sum_t \xi_{tj} + D_{jm})}\cdot$$
$$tr[(S_j - \sum_t \xi_{tj}\tilde{\Sigma}_j)\tilde{\Sigma}_{jm}^{-1}(S_j - \sum_t \xi_{tj}\tilde{\Sigma}_j)\tilde{\Sigma}_j^{-1}]$$

since the $tr[\cdot]$ part in above two equations are positive, by selecting D_j to keep $(\sum_t \xi_{tj} + D_j) > 0$, both $vec[\mu_j - \tilde{\mu}_j]^T \cdot vec[\nabla_{\mu_j} H]$ and $vec[\Sigma_j - \tilde{\Sigma}_j]^T \cdot vec[\nabla_{\Sigma_j} H]$ can ensure to be positive. Under mild conditions, the harmony functional has an upper bound, this algorithm can monotonically increase the harmony functional and converges to a local optimum.

As for the setting of D_j, it is usually set on a per-Gaussian level, e.g., a global constant E multiplied by the accumulated item $\sum_t \xi_{tj}$. when it is set a large value, training becomes slow but more stable. The relation of the proposed updating rule and the gradient can be found as:

$$\mu_j - \tilde{\mu}_j = \frac{1}{\sum_t \xi_{tj} + D_j} \tilde{\Sigma}_j \frac{\partial H}{\partial \mu_j}$$

$$\Sigma_j - \tilde{\Sigma}_j = \frac{1}{0.5(\sum_t \xi_{tj} + D_j)} \tilde{\Sigma}_j^{-1} \frac{\partial H}{\partial \Sigma_j} \tilde{\Sigma}_j^{-1}$$

As can be seen in the above equation, also by keeping $(\sum_t \xi_{tj} + D_j) > 0$, since $\frac{\partial H}{\partial \Sigma_j}$ is positive definite and $\tilde{\Sigma}_j^{-1}$ is diagonal and positive, the positive updates of the variance Σ_j can be automatically ensured in the proposed algorithm.

2.3. Integrated with Baum-Welch training algorithm

The proposed EM-like BYY-GMM algorithm can be naturally integrated into BW training framework. To be more specific, at E-step, the posterior probability γ_{tjm} (j and m denote the indices of state and Gaussian component respectively), is obtained as same as in standard EM training.

Then, ν_{tjm} can be calculated by:

$$\nu_{tjm} = (1 + \delta_{tjm})\gamma_{tjm}, \tag{9}$$

in which δ_{tjm} is calculated according to Eq. (4).

At M-step, model parameters are updated as follows:

$$\alpha_{jm} = \sum_t \nu_{tjm} / \sum_l \sum_t \nu_{tjl} \tag{10}$$

$$\mu_{jm} = \frac{\sum_t \nu_{tjm} x_t + D_{jm} \tilde{\mu}_{jm}}{\sum_t \nu_{tjm} + D_{jm}} \tag{11}$$

$$\Sigma_{jm} = \frac{\sum_t \nu_{tjm} [(x_t - \mu_{jm})(x_t - \mu_{jm})^T] + D_{jm} \tilde{\Sigma}_{jm}}{\sum_t \nu_{tjm} + D_{jm}} \tag{12}$$

The above alternation is repeated until harmony functional converges. During training, when $\alpha_{jm} \to 0$ the corresponding Gaussian component can be discarded, thus automatic model selection is achieved.

An important difference can be seen compared with standard EM-ML algorithm: ν_{tjm} actually provides a new allocation scheme. If $\delta_{tjm} > 0$, updating goes along the same direction of the ML learning but with an increased strength. If $0 > \delta_{tjm} > -1$, i.e., the fitness is worse than the average and thus this is doubtful, updating still goes along the same direction of the ML learning while with a reduced strength. When $\delta_{tjm} < -1$, updating reverses to the opposite direction, i.e., becoming de-learning (details are referred to [9]).

3. EXPERIMENTS

We carried out speech recognition experiments on Hub4 Mandarin broadcast news database to test the performance and efficiency of the proposed training algorithm. The training set is 1997 Mandarin broadcast news speech corpus (Hub-4NE) training data which consist of about 30 hours of speech. The test set is Mandarin broadcast news (Hub-4NE) evaluation data which consist of about one hour speech.

Feature extraction is Mel-cepstrum based, with corresponding first and second order time derivatives resulting in 39 dimensional features. Channel normalization is applied using cepstral mean normalization over each utterance. Acoustic models in our experiments are context dependent phoneme based acoustic models in which each unit is modeled by 3-state left-to-right HMM.After decision tree-based state tying, the baseline acoustic model totally consists of 3000 tied states.

In EM-ML training, the number of mixture components per senone is uniformly set to a constant empirically. Gaussian splitting strategy is used to increase model size from single Gaussian distribution to mixture of Gaussian distributions. First, we examine recognition performance of models trained using EM-ML in different model size. Results are given in Table 1. As can be seen from Table 1, the performance seems saturate when the mixture number increase from 16 to 32.

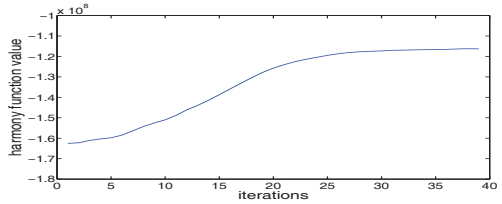
Table 1. Recognition WER function of GMM components.

# of Component	4	8	16	32	64
WER	26.12	23.17	21.98	21.89	23.29

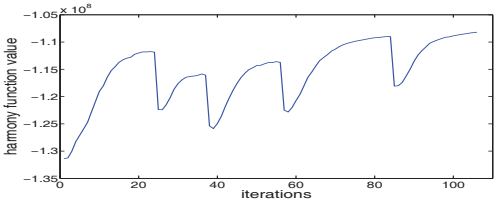
We experimented with two training strategies by using different initialization method:

In Strategy A, HMMs with each state modeled by single Gaussian are trained using EM-ML, then we simply perform five times of Gaussian splitting consecutively to increase the single Gaussian model to 32 GMMs and this model is used as the initial model to perform our proposed training algorithm, that is, the model is initialized with a large enough scale and parameters are learned and pruned during training.

In Strategy B, similar as standard EM-ML training procedure except that after each operation of Gaussian splitting, models are trained by the proposed training algorithm, that is, the model scale is increased incrementally and there is model selection performed along with different model size.



(a) The value of harmony functional in Strategy A



(b) The value of harmony functional in Strategy B

Fig. 2. The value of harmony functional in the training procedure.

In our experiments, D_{jm} is selected by principles described in last section with the constant E set to be 2.0. Training iteration is stopped when harmony functional converges. Figure 2 shows the harmony functional value in the training procedure. As can be

seen, in strategy A harmony functional converges; in strategy B, each times drop of harmony functional value correspond to an operation of Gaussian splitting.

For comparison, we also performed experiments using classical model selection criterion as the AIC and BIC. In these experiments, acoustic model with different size of GMMs are first obtained with standard ML training, then AIC and BIC are used to chose best model scale for each Gaussian mixture. Model selection and speech recognition results are given in Table 2.

Table 2. Results of models trained using different approach.

	EM-ML	BYY strategy A	BYY strategy B	BIC	AIC
Aver # of Comp	32.00	19.96	22.79	10.39	25.29
WER	21.89	20.80	21.21	23.48	22.21

As can be seen, GMM size for both models trained with strategy A and strategy B are effectively reduced, and both achieve improvements on recognition performance, by integrating BYY harmony learning into BW training, more precise model parameters are generated by model selection and less suffering from overfitting problems. Its superiority over BIC and AIC owing to that the estimating performance of BIC and AIC can deteriorates when data dimension is high and model scale is large, while in the BYY approach the extra model scale is automatically pruned during training. Another remark concerns the time consumption, the proposed training algorithm has similar convergence speed with EM-ML training, so in strategy B, the time consumption is similar with EM-ML training, while in strategy A, since models are initialized with high complexity, the overall training time is about 2 times of EM-ML training. The time consumed for BIC and AIC is nearly 1.2 times respectively of EM-ML training, since acoustic models with different model size can be preserved during the standard "split and train" procedure.

Table 3. Detailed results with testing data divided into two parts.

	EM-ML	BIC	AIC	BYY (strategy A)
SNR $>$ 20dB	16.61	17.32	16.63	14.89
SNR $<$ 20dB	28.96	31.74	29.68	29.09

To further investigate the improved performance achieved by the BYY learning, we divided the test utterances into two parts according to their SNR. Detailed results are shown in Table 3. As can be seen, for clean speech (sentences with SNR $>$ 20dB), BYY achieves significant improvements compared not only with EM-ML training but also with BIC and AIC method. For the other part of speech (sentences with SNR $<$ 20dB), the best recognition WER appears in EM-ML training. The explanation for this phenomenon should be that, interferences in speech utterances including background music, background speech and other background noise, which need some room to accommodate in order to reduce disturbances to those rooms for signals. All the three model selection methods seem not good on measuring this part of model complexity. As shown in Table 2, the EM-ML uses 32 components and thus get the best WER 28.96 while BIC uses 10.39 components get the worst WER 31.74, and AIC improves to a WER 29.68 by using 25.29 components. Interestingly, BYY uses 19.96 ($<$ 25.29 by AIC) components but still get a WER 29.09 that is better than 29.68 by AIC and is very closely to the best WER at 28.96.

4. CONCLUSION

In this paper, we first propose an EM-like BYY learning algorithm in which parameters can be updated in batch mode, then apply this algorithm into speech recognition systems to determine the number of Gaussian mixtures. Experiments are performed on Hub4 Mandarin speech data set. We compared the proposed algorithm with both standard EM-ML training and classical model selection criterions including BIC and AIC. Results show that model selection can be effectively performed by making best harmony, and also lead to improvements on recognition WER, which confirm that BYY learning suffer less from overfitting problems. In the future work, more thorough study about the selection of the D_{jm} need to be carried out. Moreover, how to improve its performance in complicated background noise environment and also the application of BYY learning for the whole GMM-HMMs to determine the model topology need to be investigated.

5. ACKNOWLEDGMENT

The work was supported in part by the National Natural Science Foundation of China-NSFC (60535030, 60605016), the National Key Basic Research Program of China-NKBRPC (2004CB318005, 2004CB318105, 2009CB825404), and the National High Technology Research and Development Program of China-NHTRDPC (2006AA010103). Lei Xu is supported by Chang Jiang Scholars Program, Chinese Ministry of Education for Chang Jiang Chair Professorship in Peking University.

6. REFERENCES

[1] Akaike, H, "Information theory and an extension of the maximum likelihood principle," *2nd International Symposium on Information Theory*, pp. 267-281, 1973

[2] S. S. Chen, P. S. Gopalakrishnan, "Clustering via the bayesian information criterion with applicationsin speech recognition," *ICASSP-98*, pp. 645-648, 1998

[3] K. Shinoda and W. Iso, "Efficient reduction of Gaussian components using MDL criterion for HMM-based speech recognition ," *ICASSP-02*, pp. 869-872, 2002

[4] F. Valente, C. Wellekens, "Variational Bayesian GMM for speech recognition," *EUROSPEECH-03*, pp. 441-444, 2003

[5] S. Watanabe, Y. Minami, A. Nakamura, N. Ueda , "Variational Bayesian estimation and clustering for speech recognition," *IEEE trans on speech and audio processing*, 2004

[6] D. Povey, P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *ICASSP-02*, pp. 105-108, 2002.

[7] P. C. Woodland, D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, pp. 25-47, 2002

[8] L. Xu, "Bayesian-Kullback coupled Ying-Yang machines: Unified learnings and new results on vector quantization," *International Conference on Neural Information Processing (ICONIP)*, pp. 977-988, 1995

[9] L. Xu, "Baysian Ying Yang learning," *in Scholarpedia 2(3):1809, http://scholarpedia.org/artical/BayesianYingYangLearning*, 2007

[10] L. Xu, "Bayesian Ying Yang System, Best Harmony Learning, and Gaussian Manifold Based Family," *In J.M. Zurada et al. (Eds.) Computational Intelligence: Research Frontiers, WCCI2008 Plenary/Invited Lectures*, LNCS5050, pp.48-78, 2008

[11] L. Xu, "Bayesian Ying-Yang System, Best Harmony Learning, and Five Action Circling," to appear in an invited special issue on *Emerging Themes on Information Theory and Bayesian Approach, Frontiers of Electrical and Electronic Engineering in China*, a journal jointly published by Higher Education Press of China and Springer, 2010