Optimization of Gaussian Mixture Model Parameters for Speaker Identification

Q.Y. Hong, Sam Kwong, and H.L. Wang

Department of Computer Science, City University of Hong Kong, Hong Kong, China
qyhong,@cs.cityu.edu.hk, {cssamk,wanghl}@cityu.edu.hk

Abstract. Gaussian mixture model (GMM) [1] has been widely used for modeling speakers. In speaker identification, one major problem is how to generate a set of GMMs for identification purposes based upon the training data. Due to the hill-climbing characteristic of the maximum likelihood (ML) method, any arbitrary estimate of the initial model parameters will usually lead to a sub-optimal model in practice. To resolve this problem, this paper proposes a hybrid training method based on the genetic algorithm (GA). It utilizes the global searching capability of the GA and combines the effectiveness of the ML method.

1 The Proposed Algorithm

Spoken utterances convey both linguistic and speaker-specific information. In the GMM-based speaker identification, the distribution of feature vectors extracted from a speaker's utterance is modeled by a weighted sum of mixture components and the speaker model that has the highest likelihood score for the utterance will be selected as the identification result. Before the identification task, the model parameters must be trained to describe the observation sequences of the speaker accurately. The traditional ML method could update the parameters repeatedly but usually only reach a local maximum. The genetic algorithm provides the global searching capability to the optimization problem. Therefore, the GMM training can escape from the initial guess and find the optimal solution if we apply the GA to the training process.

The GA method has been successfully applied for HMM-based speech recognition [2]. In this paper, we extend it to the GMM training and use the ML re-estimation as a special heuristic operator, the iteration time of which is experimentally determined to have a good balance between the searching capability and converging speed. In the proposed GA, the phenotype of a single speaker model is directly represented with the GMM structure in that the diagonal covariance is assumed. In any case, the sum of the mixture weight must satisfy the statistical constraint and will be normalized to 1.0. Moreover, the model parameters in the same mixture are actually correlated so that we form them as a unit in the crossover operator. The basic data type of the elements of the GMM is real number, so we use real number string instead of bit-string as the representation of the individuals and have the following form:

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1310–1311, 2004.

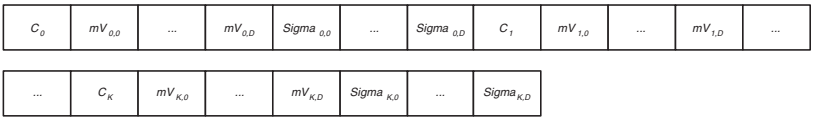


Fig. 1. The representation of the chromosome in the GA training

where C_k, $mV_{k,l}$ and $Sigma_{k,l}$ are the weight, lth mean element and lth diagonal covariance element of the kth mixture component, respectively. The fitness is defined as the average of the log-likelihood of the utterances based on the given model.

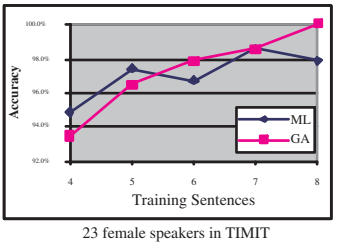
In our GA, five mixtures are randomly selected from the parent for the crossover. Mutation introduces local variations and is applied for each model parameter, which is multiplied by a Gaussian random number generator with mean = 1.0 and variance = 0.001. In each generation, the individual in the offspring is re-estimated five times by the ML operator. Each individual in the population will be further re-estimated eight times every ten GA generations.

2 Experimental Results

From the TI46 corpus, we selected 10 command words of the 8 female speakers to conduct the identification experiment. There were 100 training utterances per speaker and the whole test sections were used as the test data. The left part of Fig. 2 gives the results for different mixture number. It is seen that in all cases, the GMMs trained by our GA had higher values of fitness than the GMMs trained by the ML method. Another experiment was based on TIMIT and 23 female speakers were used. When there were 6 or more training sentences, the performance of our GA was equal to or better than the ML method. Therefore, our training approach was more preferred for sufficient training data.

Speaker	16 mixtures		32 mixtures		64 mixtures	
	ML	GA	ML	GA	ML	GA
F1	-3787.8	-3786.8	-3705.7	-3701.3	-3618.0	-3609.6
F2	-3791.2	-3779.3	-3700.2	-3691.5	-3619.9	-3604.4
F3	-4113.2	-4104.5	-4036.5	-4028.7	-3944.0	-3935.7
F4	-4162.0	-4154.7	-4083.4	-4078.1	-3988.2	-3982.0
F5	-4133.0	-4122.4	-4045.2	-4032.5	-3948.8	-3932.5
F6	-4516.0	-4512.3	-4434.0	-4423.3	-4351.7	-4336.3
F7	-4416.5	-4413.7	-4320.0	-4307.8	-4199.8	-4192.8
F8	-3874.2	-3869.8	-3790.6	-3783.5	-3689.2	-3678.6
Accuracy	89.6%	90.4%	94.2%	95.8%	97.7%	98.2%

8 female speakers in TI46



23 female speakers in TIMIT

Fig. 2. Experimental results of fitness and identification accuracy

References

1. D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17 (1995) 91-108.
2. S. Kwong, C.W. Chau, K.F. Man, and K.S. Tang. Optimisation of HMM topology and its model parameters by genetic algorithms. Pattern Recognition 34 (2001) 509-522.