The Estimating Optimal Number of Gaussian Mixtures Based on Incremental *k*-means for Speaker Identification

Younjeong Lee[1], Ki Yong Lee[2], and Joohun Lee[3]

[1,2]School of Electronic Engineering, Soongsil University, Dongjak-ku, Seoul, Korea
[3]Dept. of Internet Broadcasting, Dong-Ah Broadcasting College, Anseong, Korea

youn@ssu.ac.kr

Abstract

Gaussian mixture model (GMM) is generally used to estimate the speaker model from speech for speaker identification. In this paper, we propose the method that estimates the optimal number of Gaussian mixtures based on incremental *k*-means for speaker identification. In the proposed method, the initialization with the optimal number of mixtures is done by adding dynamically the number of mixtures one by one until the mutual relationship between any two mixtures becomes dependent. The effectiveness of the proposed method is proven by two experiments.

Keyword: Gaussian mixture model, Incremental k-means algorithm, Mutual relationship, Speaker identification.

I. Introduction

The mixture models involve the estimation of unknown parameters from a given set of observations in a variety of fields such as pattern recognition, speech and image signal analysis and static analysis[1,2] including speaker identification, the process of automatically recognizing a speaker using one's intrinsic information in his speech waves. In speaker identification, GMM has been widely used for modeling speaker's identity[3]. The parameters of the GMM for speaker models, in general, are estimated iteratively using the EM (Expectation-Maximization) algorithm, which converges to the ML estimate of the mixture parameters.

However, the EM algorithm has been known to suffer from several problems. One of the problems is that, as a local method, it is too sensitive to the selected initial parameter estimates, and it may converge to the boundary of parameter leading to inaccurate estimation[2,4]. Therefore, to solve these problems of initialization, several methods has been studied using one or a combination of those strategies such as multiple random starts and choosing the final estimation with the highest likelihood[5], and initialization by clustering algorithm[6]. Another problem is that the number of mixture which affects the overall performance of EM is not known exactly beforehand. Therefore, how to get the optimal number of mixtures for the mixture models is important to ensure an efficient and accurate estimation. To estimate the optimal number of mixtures, there have been several studies, for example, Akaike's information criterion(AIC)[7], Schwarz's Bayesian inference criterion(BIC)[8], the integrated classification likelihood (ICL) criterion[9], the mutual information(MI) with removing mixture on this subject[10] and Greedy mixture learning(GML) [11,12].

However, in those methods, to find the optimum number of mixtures, testing whether each candidate for the optimal number fits some criterion is performed for all M possible candidates ranging appropriately from M_{min} to M_{max}. This means that the final decision isn't made before every log-likelihood is calculated for all M one by one. Therefore, it brings much computational load inevitably. Also, in the conventional methods, log-likelihood is sensitive to the number of parameters to get the exact estimation of the number of mixture.

To solve these problems, we propose an efficient method that applies the mutual relationship to the initialization of EM algorithm based on incremental *k*-means so to estimate the optimal number of mixtures for GMM. First, we initialize the EM algorithm by estimating the center of mixture by using global search procedure[11]. Second, we repeat EM algorithm until likelihood converges to some appropriate value. Finally, we decide whether newly added mixture is statistically independent with other previous mixtures by measuring the mutual relationship between them. This procedure is repeated while increasing the number of mixtures one by one until at least one of their mutual relationship turns out positive. In our method, since the repeating procedure stops right after the above criterion is satisfied even though the test gets to the last candidate M_{max} yet, it saves much computational load compared to the conventional methods. Also, with the accurate initialization, high performance of parameter estimation can be achieved.

The remaining part of this paper is organized as follows. In section 2, we describe the mixture model with the initialization of the EM algorithm for the maximum likelihood training and explain a method to investigate the mutual relationship of components to estimate the optimal number of mixtures. In section 3, we describe the algorithm of the proposed method. The section 4 gives the method of speaker identification. The experimental results are also summarized in the section 5. In section 6, we draw conclusions.

II. The Estimating Optimal Number of Gaussian Mixtures Based on Incremental *k*-means

A. *The initialization of EM based on Incremental k-means algorithm for GMM*

Suppose we have T observations $X = \{x_1, \cdots, x_T\},\ x_t \in R^d$. A Gaussian mixture density is defined by a weighted linear combination of M component densities as

$$p(x_t \mid \theta) = \sum_{i=1}^{M} w_i b_i(x_t) \tag{1}$$

where the component densities, $b_i(x_t)$, are defined by *d*-multivariate Gaussian function of the form

$$b_i(x_t) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1} (x_t - \mu_i) \right\} \tag{2}$$

with mean vector μ_i and covariance matrix Σ_i for i-th mixture. The mixture weights w_i, furthermore satisfy the constraint $\sum_1^M w_i = 1$. Collectively, the parameters of speaker's density model are denoted as $\theta = \{w_i, \mu_i, \Sigma_i\}_{i=1}^M$ [3].

To estimate parameters from all sequences of T vectors X, the GMM likelihood can be written as

$$P(X \mid \theta) = \prod_{t=1}^{T} p(x_t \mid \theta) \tag{3}$$

In general, the ML estimation is used to estimate the parameters θ which maximize the likelihood of the GMM. Since the GMM likelihood of the nonlinear function is impossible that maximizes directly, the ML estimations can be possible by using the EM algorithm iteratively.

Since the EM algorithm depends on initialization, we use the method that is modified using the initialization based on incremental k-means algorithm. The EM algorithm with the initialization estimates a new model by alternating two steps until likelihood is converged[3, 4]. Assume auxiliary Q-function is followed by

$$Q(\theta, \hat{\theta}) = \sum_{t=1}^{T} \sum_{i=1}^{M} p(i \mid x_t, \hat{\theta}) \{\log w_i + \log b_i(x_t)\} \tag{4}$$

Initialization of mixtures: The EM algorithm can be initialized by the obtained μ_k, $k = 1, \cdots, M$ of M mixtures from given data, and converge rapidly and well for ML estimate of models. Each observed data x_t is used for a candidate center of mixtures of the GMM. And then, a center of mixture, x_n, is selected from a criterion which is defined by the exponential function of Euclidean norm to all of the other data.

$$E_n^1 = \sum_{t=1}^{T} \exp\left\{-\|x_t - x_n\|^2\right\} \quad , \quad 1 \le n \le T,\ k=1 \tag{5}$$

Obviously, the first center of the mixture has the highest density among all of data when it is surrounded by many neighboring data and can be selected as

$$\mu_1 = x_{n^*}, \quad \text{where} \quad n^* = \arg\max_{1 \le n \le T} E_n^1 \ . \tag{6}$$

Given the (k-1)-th mixture, for k-th the criterion, we can modify the criterion of each data by the following equation:

$$E_n^k = \sum_{t=1}^{T} \left\{ \sum_{m=1}^{k-1} I(x_t \in C_m) \|x_n - \mu_m\|^2 + I(x_t \in C_k) \|x_t - x_n\|^2 \right\}, \quad 2 \le k \le M \tag{7}$$

where $1 \le n \le T$, C_k is k-th component of M mixtures and $I(X) = 1$ if X is true and 0 otherwise [11]. Using distance criterion, we can find the mean of the k-th mixture with the minimum criterion.

$$\mu_k = x_{n^*}, \quad \text{where} \quad n^* = \arg\min_{1 \le n \le T} E_n^k \tag{8}$$

E-Step: In the E-step, we consider the conditional expectation of the log-likelihood, given X and the current estimate $\hat{\theta}$. The E-step of the EM algorithm involves the use of the current parameter estimates to assess the proportions of the mixture components that is called a posteriori probability $\zeta_{it} = p(i \mid x_t, \hat{\theta})$. Applying Bayes' theorem, the posteriori probability of x_t for *i*-th mixture is defined by

$$\zeta_{it} = p(i \mid x_t, \hat{\theta}) = \frac{p_i b_i(x_t)}{\sum_{h=1}^{M} p_h b_h(x_t)} \tag{9}$$

Given the mean μ_k of mixture, which is decided by the data point x_{n^*}, the parameters is reestimated by the EM algorithm.

M-Step: The M-step is a computation for the global maximization of $Q(\theta, \hat{\theta})$. Hence, reestimate the parameters {mixture weight, covariance matrix} with a monotone increasing likelihood according to

Mixture Weight: $$p_i = \frac{1}{T} \sum_{t=1}^{T} \zeta_{it} \tag{10}$$

Variances: $$\Sigma_i = \frac{\sum_{t=1}^{T} \zeta_{it} x_t^2}{\sum_{t=1}^{T} \zeta_{it}} - \mu_i^2 \tag{11}$$

From EM algorithm with the initialization based on incremental *k*-means for the GMM, we can compute iteratively the parameters of mixture models from the observed data.

B. *Estimation of the optimal Mixtures number*

The parameters, with too many or few components, may inappropriately estimate the mixture model. Therefore, the number of mixtures is important for the EM algorithm to the well fitted mixture model. Unlike the previous method using model selection criterion, in this section, we determine the optimal number of components as measuring mutual information between two components when the mixture based on incremental *k*-means is added. The mutual relationship is employed to investigate whether components are statistically dependent. That is, the mutual relationship measures the information shared between two components [10].

When one component is added at a time, if the relationship between the new *k*-th component and the previous components is still independent, the parameters obtaining from the *k* components regard as reestimated parameters. On the contrary, if at least one component is dependent, we should consider the parameters in the previous $(k-1)$ mixture models and determine the optimal number of mixture is $(k-1)$ at last.

The mutual relationship of components between *i* and *k* is defined as

$$\varphi(i,k) = p(i,k) \log_2 \frac{p(i,k)}{p(i)\, p(k)}, \; i = 1, \cdots, k \tag{12}$$

where $p(i)$ is the probability of the mixture i and $p(i,k)$ is the joint probability of i and k as following;

$$p(i) = \frac{1}{T}\sum_{t=1}^{T} p(i \mid x_t), \tag{13}$$

$$p(i,k) = \frac{1}{T}\sum_{t=1}^{T} p(i|x_t)p(k|x_t) \tag{14}$$

The mutual relationship is three possible values: negative, zero, and positive. If $\varphi(i,k)$ is a zero value, it means that mixture *i* and *k* are statistically independent: $p(i,k) = p(i)p(k)$. If $\varphi(i,k)$ has a positive value, it means *i* and *k* are statistically dependent. A negative value of mutual relationship means that *i* and *k* can be regarded as much less dependent. When $\varphi(i,k)$ is positive, the *k*-th component can be removed without damage to the estimated distributions. In order to obtain the optimal number of mixture, we maintain the iterative procedures until the mutual relationship between components is firstly dependent.

III. The Algorithm of Estimating the Optimal Mixtures

From a given set of observations, we need to determine the EM with initialization based on incremental *k*-means for GMM and the optimal number of mixtures. Whenever the mixture by the incremental *k*-means algorithm is newly added, we consider the mutual relationship to know whether components are statistically dependent or not. If at least two components are statistically dependent, we should stop performing the incremental number of mixtures. After then, we can conclude the previous number of mixtures ($M=k-1$). The proposed method is followed by:

Step1: For $n = 1,\cdots,T$, let *k*=1 and calculate the distance error E_n^1 and determine the center μ_1 of mixture.
Step 2: To add a component, search for center of another mixture with minimum E_n^k based on the incremental *k*-means algorithm (*k*=*k*+1).
Step 3: Given the initial models of *k*, for the reestimation of parameters, perform alternatively the EM algorithm using eq.(9), eq.(10) and eq.(11) of mixture model until likelihood is converging. Then, the quantities, $p(i)$, $p(k)$, and $p(i,k)$ are available to calculate the mutual relationship.
Step 4: For $i = 1,\cdots,k$, calculate the mutual relationship between components by (12). If $\varphi(i,k) > 0$, then stop and set the optimal number of mixture as $M = k-1$ else return to Step 2.

If the mutual relationship of components is positive at each stage, we are able to know the optimal number of components by (*M*=*k*-1) returning to the previous stage.

IV. Speaker Identification

For speaker identification, each of S speakers is represented by GMMs, $\theta_1,\cdots,\theta_S$, respectively[3]. The object of speaker identification is to find the speaker model which has the maximum a posteriori probability for a given feature sequence as

$$\hat{s} = \arg\max_{1\le l\le S} \sum_{t=1}^{T} \log p\left(x_t \mid \theta_l\right). \tag{15}$$

V. Experimental Results

To show effectiveness of the proposed method, we performed two experiments in which artificial data set and real speech data are used.

Firstly, we performed the experiments with artificial data generated as dataset of 2000 points with two-dimensional three mixtures of Gaussian normal distribution. The correlated Gaussian normal distribution is given as

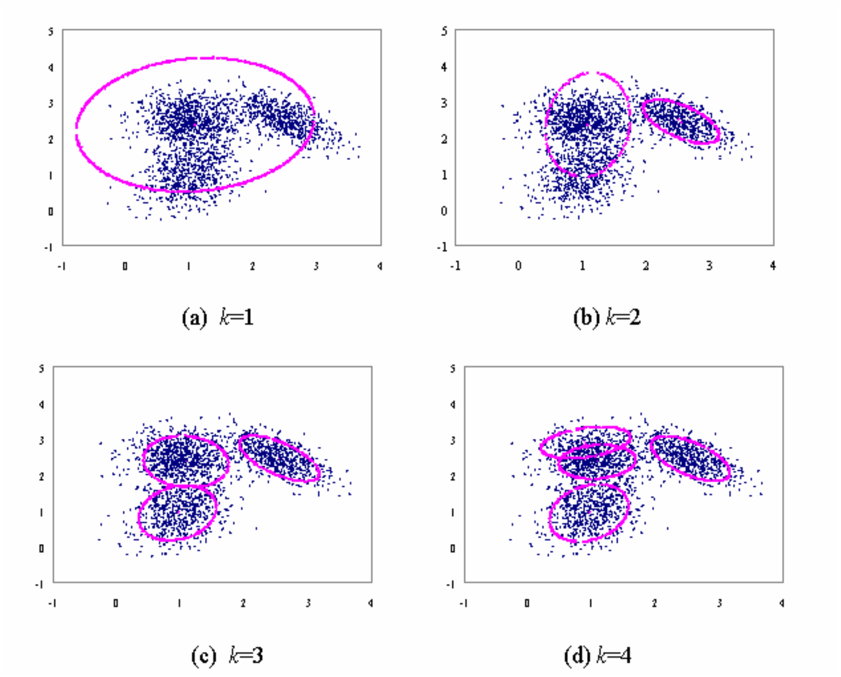
$$0.3N\left[x\left|\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}0.15 & 0.05\\0.05 & 0.25\end{pmatrix}\right.\right]+0.4N\left[x\left|\begin{pmatrix}1\\1.25\end{pmatrix},\begin{pmatrix}0.15 & 0\\0 & 0.15\end{pmatrix}\right.\right]+0.3N\left[x\left|\begin{pmatrix}2.5\\2.5\end{pmatrix},\begin{pmatrix}0.15 & -0.1\\-0.1 & 0.15\end{pmatrix}\right.\right]. \tag{16}$$

which is used by Cheung in[13]. With these artificial data, one component (i.e, mixture) is firstly obtained by the global searching and a new mixture is added to it one by one until the final number of mixtures (in this case, $k=4$) are found, i.e., at least two mixtures of them start to have the positive mutual relationship.

In table 1, the mixtures become statistically dependent at fourth step ($k=4$) since the 1st and 4th mixtures of them have positive mutual relationship (0.0266). Hence, the 4th component is removed and the previous number of mixtures (M=3) is determined as the optimal number of mixtures.

Table 1. The mutual relationship between mixtures

# of mixture(*k*)	Mean of the *i*-th Mixture	***i***	$\varphi(i,k)$
1	(1.0788,2.4606)	1	~
2	(2.5416,2.4477)	1	-0.0611
3	(0.9467,0.9378)	1	-0.0603
		2	-0.0017
4	(0.9013,2.8942)	1	0.0266
		2	-0.0170
		3	-0.0060



(a) k=1

(b) k=2

(c) k=3

(d) k=4

Fig. 1. The construction of a three-component mixture distribution

An example is given in figure 1, which depicts the evolution of a solution for above artificially generated data. The mixtures $k_1,\ldots k_4$ are depicted; each component is shown as an ellipse that has the covariance matrix as axes and radii of twice the square root of the corresponding eigenvalue. In fig 1-(a), the mixture starts with the number of mixture being set to one. According to the increased number of mixture, the mixture changes as shown in the following fig 1-(b), (c) and (d). Accordingly, in figure 1-(d), since the relationship between the first mixture and the fourth one has a positive value which means that they are statistically dependent; the optimal number of mixture is set to three, which is the previous step before the final one. This means that three is the optimum number guaranteeing all mixtures are statistically independent respectively.

Secondly, we performed the speaker identification using real speech data. The speech data consists of a Korean sentence uttered 15 times each by 100 females and 100 males. Speaker models were trained using 10 utterances of each speaker. The remaining 5 utterances were used for test of speaker identification. The speech data was sampled at 16 kHz and was parameterized using 12 MFCC and 13 delta cepstrum. The analysis window size was 20ms with 10ms overlap. For speaker identification, at first we used the original GMM with fixed number of mixtures (5~30) and the conventional methods including AIC, BIC, ICL, MI and GML. In the methods of estimating the optimal mixtures, the number of mixtures is obtained from the average of the estimated numbers of all speakers. And then, to compare those to our method, we performed the same speaker identification with the optimal number of mixtures, which is found by the proposed method for each speaker model.

Table 2 show performances of the speaker identification in those conventional methods and our method. The previous methods show the following results: AIC method with 28 mixtures has 95.5%, BIC method with 21 mixtures has 96.9%, ICL method with 20 mixtures has 98.1%, and the decremental method by MI with 20 mixtures has 97.7%. Finally, the GML method with 19 mixtures has 98.6% and the proposed method shows higher performance as 98.8% with 18 mixtures.

Table 2. The performance of speaker identification

Algorithm	Avg. Num. of mixtures	Avg. Processing Time(Sec)	Identification Rates (%)
AIC	28	21.02	95.5
BIC	21	15.25	96.9
ICL	20	15.33	98.1
MI	20	2.05	98.7
GML	19	6.32	98.6
The proposed method	18	1.99	98.8

In addition to having the highest performance, from the viewpoint of complexity, our method is superior to others since our method avoids examining further the rest number of mixtures once it gets the optimal number of mixtures while others have to inevitably examine all candidate number of mixtures. Also, the proposed method shows the highest performance as 98.8% with 18 mixtures. Moreover, the average processing time during a speaker training of the proposed one is the fastest. The number of mixtures is the value obtained from averaged one of the estimated numbers of all speakers.

VI. Conclusion

In this paper, we proposed the method of the estimating number of Gaussian mixtures based on incremental *k*-means algorithm for speaker identification. The proposed method is initialized by using incremental *k*-means to guarantee the global convergence of EM algorithm. By using incremental *k*-means, the optimal number of mixture is searched for candidate numbers by increasing the number one by one until appropriate criterion is satisfied. By applying the optimum number of mixture and the initial values to EM, we can get more accurate parameter estimation. When applying our method to speaker identification, we can obtain higher performance than the typical methods. Additionally, since our method does not examine all candidates while others do, it saves much computational load. The experimental results show that the proposed method is an effective and fast algorithm to find accurate parameters with obtaining the optimal number of mixtures for GMM.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

References

[1] P.Paclik, J.Novovičová, “Number of components and initialization in Gaussian Mixture Model for pattern recognition,” *In Proceedings of the 14th ICPR*, *Australia,* 1998, pp. 886-890.

[2] M.A.T.Figueiredo, A.K. Jain, “Unsupervised Learning of finite mixture models,” *IEEE Trans. on PAMI.*, vol. 24, no.3, 2002, pp. 381-396.

[3] D.A. Reynolds and R.Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. on SAP*, vol.3, no. 1, 1995, pp. 72-82.

[4] A.Dempster, N.Laird, and D.Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J.Roy.Statist. Soc. Ser.,* B39, 1977, pp. 1-38.

[5] S.Richardson and P.Green, “Bayesian Approaches to Gaussian Mixture Modeling,” *IEEE Trans. on PAMI* , vol.2, 1997, pp. 243-252.

[6] G.Mclachlan and D. Peel, *Finite Mixture Models*, New York: John Wiley & Sons, 2000.

[7] H.Akaike, “Information theory and an extension of the maximum likelihood principle,” *In second International Symposium on Information Theory*, eds. V.N. Petrov and F. Csaki, Budapest: Akailseoniai-Kiudo, 1973, pp. 267 -281.

[8] G.Schwarz, “Estimating the Dimension of a Model,” *Annals of Statistics*, vol. 6, 1978, pp. 461-464.

[9] C. Biernacki, G.Celeux and G. Govarert, “Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood,” Technical Report 3,521, Inria, 1998.

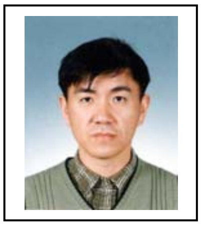
[10] Z.R.Yang, and M.Zwolinski, “Mutual information theory for adaptive mixture models,” *IEEE Trans. on PAMI* , vol.23, no.4, 2001, pp. 396-403.

[11] A.Likas, N.Vlassis, and J.Verbeek, “The Global *k*-means clustering algorithm,” *Pattern Recognition* 36, 2003, pp. 451-461.

[12] J.Verbeek, N.Vlassis, and B.Krose, “Efficient Greedy Learning of Gaussian Mixture,” *Neural Computation* 15, 2003, pp. 469-485.

[13] Y.Cheung, “k^*-means: A new generalized *k*-means clustering algorithm,” *Pattern Recognition Letters 24,* 2003, pp.2883-2893.

Younjeong Lee received the Ph.D. degree in Electronic Engineering from Soongsil University in 2006. Her research interests include the biometrics and its application as well as speech signal processing.



Ki Yong Lee was a professor in the School of Electronic Engineering, Soongsil University. He received the Ph.D. degree in Electronic Engineering from Seoul National University in 1991. His research interests include the speaker recognition and its application as well as speech enhancement.

Joohun Lee is an associate professor in the Department of Internet Broadcasting, Dong-ah broadcasting college. He received the Ph.D. degree in Electronic Engineering from Seoul National University in 1995. His research interests include the biometrics and its application as well as broadband TV system.