MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Effective and Efficient Sports Highlights Extraction Using the Minimum Description Length Criterion in Selecting GMM Structures

Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, Thomas S. Huang

TR2004-061 June 2004

Abstract

In fitting the training data with Guassian Mixture Models (GMMs) or appropriate structures using the MDL criterion, we are able to improve audio classification accuracy with a large margin. With the MDL-GMMs, we are also able to greatly improve the accuracy in extracting sports highlights. Since we have focused on audio domain processing, it enables us to extract highlights very fast. In this paper, we have demonstrated the importance of a better understanding of model structures in such a pattern recognition task.

ICME 2004

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2004
201 Broadway, Cambridge, Massachusetts 02139

EFFECTIVE AND EFFICIENT SPORTS HIGHLIGHTS EXTRACTION USING THE MINIMUM DESCRIPTION LENGTH CRITERION IN SELECTING GMM STRUCTURES

Ziyou Xiong†, Regunathan Radhakrishnan‡, Ajay Divakaran‡ and Thomas S. Huang†

†Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, Urbana, IL, USA
‡Mitsubishi Electric Research Laboratories, Cambridge, MA, USA
E-mail: {zxiong, huang}@ifp.uiuc.edu, {regu, ajayd}@merl.com

ABSTRACT

In fitting the training data with Guassian Mixture Models(GMMs) of appropriate structures using the MDL criterion, we are able to improve audio classification accuracy with a large margin. With the MDL-GMMs, we are also able to greatly improve the accuracy in extracting sports highlights. Since we have focused on audio domain processing, it enables us to extract highlights very fast. In this paper, we have demonstrated the importance of a better understanding of model structures in such a pattern recognition task.

Keywords*: Sports Highlights Extraction, Model Structure, Gaussian Mixture Models, Minimum Description Length*

1. INTRODUCTION AND RELATED WORK

Sports highlights extraction is one of the most important applications of video analysis. Approaches based on audio classification[1], video feature extraction and highlights modelling[2] have been reported. Due to the space limitation, please refer to the introduction section of [1][2][3] [4] for a detailed literature survey.

We have reported our research results that are built upon a foundation of audio classification framework[3][4]. This paper presents an approach that makes considerable improvement on that foundation. It is motivated by finding a solution to the following shortcoming of the GMMs. Traditionally the GMMs are assumed to have the same number of mixtures for a classification task. This single, "optimal" number of mixtures is usually chosen through cross validation. The practical problem is that for some class this number will lead to over-fitting of the training data if it is much less than the actual one or inversely, under-fitting of the data. Our solution is to use the MDL criterion in selecting the number of mixtures. MDL-GMMs fit the training data to the generative process as closely as possible, avoiding the problem of over-fitting or under-fitting.

2. ESTIMATING THE NUMBER OF MIXTURES IN GMMS USING MDL

2.1. Theoretical Derivations

The derivations here follow those in [5]. Let Y be an M dimensional random vector to be modelled using a Gaussian mixture distribution. Let K denote the number of Gaussian mixtures, and we use the notation π, μ, and R to denote the parameter sets $\{\pi_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$ and $\{R_k\}_{k=1}^K$ for mixture coefficients, means and variances. The complete set of parameters are then given by K and $\theta = (\pi, \mu, R)$. The log of the probability of the entire sequence $Y = \{Y_n\}_{n=1}^N$ is then given by

$$\log p_y(y|K,\theta) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} p_{y_n|x_n}(y_n|k,\theta)\pi_k \right) . \quad (1)$$

The objective is then to estimate the parameters K and $\theta \in \Omega^{(K)}$. The maximum likelihood (ML) estimate is given by

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Omega^{(K)}} \log p_y(y|K,\theta)$$

the estimate of K is based on the minimization of the expression

$$MDL(K,\theta) = -\log p_y(y|K,\theta) + \frac{1}{2} L \log(NM) , \quad (2)$$

where L is the number of continuously valued real numbers required to specify the parameter θ. In this application,

$$L = K\left(1 + M + \frac{(M+1)M}{2}\right) - 1 .$$

Notice that this criterion has a penalty term on the total number of data values NM, suggested by Rissanen [6] called the minimum description length (MDL) estimator. Let us denote the parameter learning of GMMs using the MDL criterion MDL-GMM.

While the Expectation Maximization(EM) algorithm can be used to update the parameter θ, it does not provide a solution to the problem of how to change the model order K. Our approach will be to start with a large number of clusters, and then sequentially decrement the value of K. For each value of K, we will apply the EM update until we converge to a local minimum of the MDL functional. After we have done this for each value of K, we may simply select the value of K and corresponding parameters that resulted in the smallest value of the MDL criterion.

The question remains of how to decrement the number of clusters from K to $K-1$. We will do this by merging two closest clusters to form a single cluster. More specifically, the two clusters l and m are specified as a single cluster (l,m) with prior probability, mean and covariance given by

$$\pi^*_{(l,m)} = \bar{\pi}_l + \bar{\pi}_m \tag{3}$$

$$\mu^*_{(l,m)} = \frac{\bar{\pi}_l\bar{\mu}_l + \bar{\pi}_m\bar{\mu}_m}{\bar{\pi}_l + \bar{\pi}_m} \tag{4}$$

$$R^*_{(l,m)} = \frac{\bar{\pi}_l\left(\bar{R}_l + (\bar{\mu}_l - \mu_{(l,m)})(\bar{\mu}_l - \mu_{(l,m)})^t\right)}{\bar{\pi}_l + \bar{\pi}_m} + \frac{\bar{\pi}_m\left(\bar{R}_m + (\bar{\mu}_m - \mu_{(l,m)})(\bar{\mu}_m - \mu_{(l,m)})^t\right)}{\bar{\pi}_l + \bar{\pi}_m}. \tag{5}$$

Here the $\bar{\pi}$, $\bar{\mu}$, and $\bar{R}$ are given by the EM update of the two individual mixtures before they are merged.

2.2. An Example: MDL-GMM for Different Sound Classes

We've collected 679 audio clips from TV broadcasting of golf, baseball and soccer games. This database is a subset of that in [3]. Each of them is hand-labelled into one of the five classes as ground truth: applause, cheering, music, speech, "speech with music". Their corresponding numbers of clips are 105, 82, 185, 168, 139. Their duration differs from around 1 second to more than 10 seconds. The total duration is approximately 1 hour and 12 minutes. The audio signals are all mono-channel with a sampling rate of 16kHz.

We extract 100 13-dimensional Mel-Frequency Cepstrum Coefficients(MFCC) per second using a 25 msec window. We also add the first- and second-order time derivatives to the basic MFCC parameters in order to enhance performance. For more details on MFCC feature extraction, please see [7].

For each class of sound data, we first assign a relative large number of mixtures to K, calculate the MDL score $MDL(K,\theta)$ using all the training sound files, then merge the two nearest Gaussian components to get the next MDL score $MDL(K-1,\theta)$, then iterate till $K=1$. The "optimal" number K is chosen as the one that gives the minimum of the MDL scores. For the training database we have, the relationship between $MDL(K,\theta)$ and K for all five classes are shown in Fig. 1.

From Fig. 1 we observe that the optimal mixture numbers of the above five audio classes are 2, 2, 4, 18, 8 respectively. This observation can be intuitively interpreted as follows. Applause or cheering has a relatively simpler spectral structure, hence fewer Gaussian components can model the data well. In comparison, speech has a much more complex, variant spectral distribution, it needs much more components.

Also, we observe that the complexity of music is between that of applause or cheering and speech. For "speech with music", i.e., a mixture class of speech and music, its complexity is between the two classed that are in the mixture.

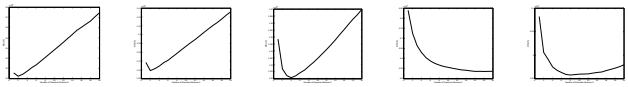

Fig. 1. $MDL(K,\theta)$(Y axis) with respect to different number of GMM mixtures K(X axis) to model Applause, Cheering, Music, Speech and "SpeechWithMusic" sound shown in the raster-scan order. $K = 1\cdots 20$. The optimal mixture numbers at the lowest positions of the curves are 2, 2, 4, 18, 8 respectively.

3. PERFORMANCE COMPARISON BETWEEN TRADITIONAL GMMS AND MDL-GMMS

To compare the two GMMs, we cross-validate the classification results by dividing the above-mentioned 5-class audio dataset into 90%/10% training/test sets. For one, the number of Gaussian mixtures is assumed to be 10 for all the classes, the test results are put into Table 1. For the other, the number of mixtures are those chosen from Fig. 1, the results are shown in Table 2. Note that the overall classification accuracy has been improved by more than 8%.

	[1]	[2]	[3]	[4]	[5]
[1]	88.8%	5.0%	3%	2%	1.2%
[2]	5%	90.1%	2%	0	2.9%
[3]	5.6%	0	88.9%	5.6%	0
[4]	0	0	0	94.1%	5.9%
[5]	0	0	6.9%	5.1%	88%
	Average Recognition Rate: 90.0%				

Table 1. Performance of traditional GMM, every class is modelled using 10 Gaussian mixtures. [1]: applause; [2] cheering; [3] music; [4] speech; [5] "speech with music". Classification accuracy on the 10% data by models trained on the 90% data.

	[1]	[2]	[3]	[4]	[5]
[1]	97.1%	0	0	0.9%	2.0%
[2]	0	99.0%	1.0%	0	0
[3]	0	1.0%	99.0%	0	0
[4]	0	0	0	99.0%	1.0%
[5]	0	0	1.0%	0	99.0%
	Average Recognition Rate: 98.6%				

Table 2. Performance of MDL-GMM: classification accuracy on the 10% data by models trained on the 90% data.

4. EXPERIMENTAL RESULTS ON SPORTS HIGHLIGHTS GENERATION

In [3], we have reported some results of sports highlights extraction based on audio classification and the correlation between applause/cheering sound with exciting moments. However, there we have not used the MDL criterion to select the model structures, so we have been using the learned models in a "blind" sense. Now equipped with the MDL-GMMs and with the observation that they can greatly improve classification accuracy, we revisit the problem in [3].

First, instead of training on 90% and testing on 10% of the data as in Table 2, we train the MDL-GMMs on all the data in the ground truth set. In order to gain a better understanding of classification, especially on the applause/cheering sound, we test also on all the data in the ground truth set before we test the game data. The results are organized into Table 3 and Table 4. The classification accuracy on either applause or cheering has been quite high.

	[1]	[2]	[3]	[4]	[5]
[1]	97.1%	0	0	0.9%	2.0%
[2]	0	99.0%	1.0%	0	0
[3]	1.0%	8.0%	89.0%	0	2.0%
[4]	0	0	0	92.2%	7.8%
[5]	0	0	0.7%	2.8%	96.5%
	Average Recognition Rate: 94.76%				

Table 3. The confusion matrix on ALL the audio data. The results are based on MDL-GMMs with different "optimal" number of mixtures(see Fig. 1).

We ran audio classification on the audio sound track of a 3-hour golf game. The game took place on a rainy day so the existence of the sound of raining has corrupted our previous classification results in [3] to a great degree. Every second of the game audio is classified into one of the 5 classes. Those contiguous applause segments are sorted according to the duration of contiguity. The distribution of these contiguous applause segments is shown in Table 5. Note that the applause segments can be as long as 9 continuous seconds.

Based on when the beginning of applause or cheering is, we choose to include a certain number of seconds of video before the beginning moment to include the play action(golf swing, par etc.), then we compare these segments to those ground-truth highlights that are labelled by human viewers.

4.1. Performance and Comparison with Results in [4] in Terms of Precision-Recall Curves

We analyze the extracted highlights that are based on those segments in Table 5. For each length L of the contiguous applause segments, we calculate the precision and recall values. (Precision is the percentage of highlights that are correct of all those extracted. Recall is the percentage of highlights that are in the ground-truth set.) We then plot the precision vs. recall values for all different L into the left of Fig. 2.

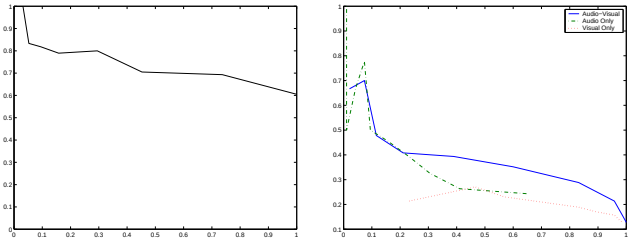

Fig. 2. Precision-recall curves for the test golf game. Left: by the current approach; Right: by the approaches in [4]. Y-axis: precision, X-axis: recall.

In comparison, the right-hand side of Fig. 2 shows the results reported in [4] on the same game, where dash-line curve shows the precision-recall relationship when a Hidden Markov Model(HMM) is used to model the highlights using the audio classes labelled by the models in Table 1. The intention there to use the HMM on top of the GMM is to enhance performance. The solid line curve shows the results when a coupled HMM is used to model both audio and video classes in order to further enhance performance

	[1]	[2]	[3]	[4]	[5]
[1]	102	0	0	1	2
[2]	0	81	1	0	0
[3]	2	15	164	0	4
[4]	0	0	0	155	13
[5]	0	0	1	4	134

Table 4. Detailed recognition results of those in Table 3. A distribution of the number of sound examples that are correctly or incorrectly classified for each class is shown.

on the dash-line curve. Although we have established the argument on the superiority of coupled HMM over audio-only HMM or video-only HMM(the dotted curve), overall the performances there are not satisfactory as the best one(coupled HMM) has poor performance at the right-most part of the curve.

From the two figures in Fig. 2, we observe that the MDL-GMMs out-perform those approaches in [4] by a large margin. For example, at 90% recall, the left-hand figure shows $\geq$ 70% precision rate, while the right-hand figure shows only $\sim$ 30% precision rate, suggesting that the false alarm rate is much lower using the current approach.

4.2. System Interface

Our system interface is shown in Fig. 3. The spiky curve at the lower half of the figure is a plot of confidence level with respect to time(second by second). Larger confidence level values indicate more likely there are highlights at the time instance. We provide a moving threshold for the user to place on the curve. Those segments with values greater than the threshold are played one after another.

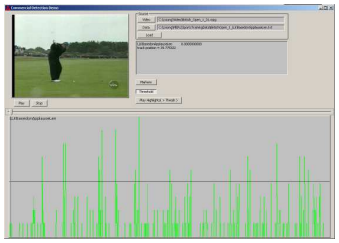

Fig. 3. The interface of our system displaying sports highlights. The horizontal line imposed on the curve is the threshold value the user can choose to display those segments with confidence level greater than the threshold.

applause length	# of instances	# of highlights
$L \geq 9s$	1	1
$L \geq 8s$	3	3
$L \geq 7s$	6	5
$L \geq 6s$	11	9
$L \geq 5s$	19	15
$L \geq 4s$	35	28
$L \geq 3s$	61	43
$L \geq 2s$	101	70
$L \geq 1s$	255	95

Table 5. Number of contiguous applause segments and highlights found by the MDL-GMMs in the golf game. These highlights are in the vicinity of the applause segments. These numbers are plotted in the left of Fig. 2.

5. CONCLUSIONS AND FUTURE WORK

We have demonstrated the importance of a better understanding of model structures in the audio analysis for sports highlights generation. In fact we looked into model parameter selection in [4] in terms of number of states of the HMMs, coupled HMMs. We showed there that the selection also improved recognition accuracy. What we report in this paper is complementary to that in the following sense: MDL-GMMs can find better GMM structures; the techniques in [4] can find better HMM structures. In the future, we will incorporate MDL-GMMs into the system in [4] to further improve the performance.

6. REFERENCES

[1] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *Eighth ACM International Conference on Multimedia*, pp. 105 – 115, 2000.

[2] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing, (ICASSP-2002)*, May 2002, Orlando, FL, USA.

[3] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Audio-based highlights extraction from baseball, golf and soccer games in a unified framework," *Proceedings. Intl. Conf. on Acoustic, Speech and Signal Processing(ICASSP)*, vol. 5, pp. 628 – 631, 2003.

[4] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Audio-visual sports highlights extraction using coupled hidden markov models," *submitted to Pattern Analysis and Application Journal*, 2004, Special Issue on Video Based Event Detection.

[5] C. A. Bouman, "CLUSTER: An unsupervised algorithm for modeling gaussian mixtures," *http://www.ece.purdue.edu/∼bouman*, School of Electrical Engineering, Purdue University.

[6] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, vol. 11, no. 2, pp. 417 – 431, 1983.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book version 3.2*, Cambridge University Press, 2003, Cambridge University Engineering Department.