Audio Data Retrieval and Recognition Using Model Selection Criterion

Konstantin Biatov
Fraunhofer IAIS
biatov@iais.fraunhofer.de

Abstract

Model selection criterion is an unsupervised technique and can be used to compare statistical distribution of the data. In this paper the experiments of using model selection criterion for audio analysis tasks are presented. This technique is applied for direct audio search in German broadcasts news with the high variability in duration and loudness of the search patterns. Using model selection criterion as a distance metric the experiments for identification of 14 environmental sounds are carried out. For environment sounds detection the decision is based on mutual similarity of compared events to the set of reference events. For audio events recognition Latent Semantic Indexing (LSI) is also tested.. Approximately 500 audio segments from 14 sound types are used in the recognition test. The experiments show that the applications of model selection criterion for direct audio search, unsupervised environmental sounds analysis and sounds recognition using LSI are effective and accurate.

1. Introduction

The size of available multimedia data is increased. It is important to have effective and reliable methods for retrieval and indexing of multimedia data. The retrieval task of non-speech audio could be divided in to two parts. The first task is to find audio segments that have an exact match with cue audio query. The examples of the first retrieval task are audio fingerprinting and jingle detection. In the last decade some methods are suggested [1], [2], [3], [4], [5]. For audio search the following overall measures between query and tested audio segment are used: spectral flatness, spectral centroids, histograms, sinusoidal peaks and etc. All of the described measures are developed to an find exact match between audio query and unknown audio data.

The second part of the retrieval task is audio events recognition and retrieval. The goal of the audio events recognition is to find audio data that corresponds to the same audio class. For audio events recognition the model based approach is usually used.

This paper concerns the technique that is efficient and accurate for exact audio match as well as for audio class detection and identification. The model selection criterion refers to choosing the most appropriate model to express the given data. Starting from Akaike criterion [6] some model selection criteria have been introduced in the last three decades [7], [8]. The model selection criteria are successfully used in variety applications in speech technology, first of all in speaker segmentation and clustering [9]. The model selection criteria takes into account the sizes of compared data that is important in the case of temporal variation of the data. This method can be considered as a method of unsupervised recognition. One of the well known model selection criterion is Bayesian Information Criterion (BIC). The purpose of this paper is to present the application of BIC for exact audio query search and for the recognition of some audio classes, in particular, for recognition of 14 environmental audio events such as airplanes, applause, car motors, car accidents, bar/restaurants, laughter, traffic, car races, town, casino, horses, weather, steps, crowds and explosions.

The BIC based direct search method is described in section 2. The data used for experiments for exact audio search are described briefly in the section 3, the direct audio search experiments are presented in the section 4, BIC based technique for unsupervised events recognition, used data and experiments are presented in the section 5, LSI based events recognition is presented in the section 6 and finally the conclusion is offered in the section 7.

2. BIC for audio search

The BIC for audio application was initially proposed in [9]. In general the BIC is defined as

$$BIC(M) = \log L(X, M) - \lambda \frac{\#(M)}{2} \log(N) \qquad (1)$$

where log L(X,M) denotes segment X likelihood given by the model M, N is the number of the feature vectors in the data, #(M) is the number of the free parameters in the model and λ is a tuning parameter. For Gaussian distributions in order to estimate the data distribution turn point between two segments c_i and c_j that have n_i and n_j frames respectively, the ΔBIC value is computed as:

$$\frac{1}{2} n_i \log|\Sigma_i| + \frac{1}{2} n_j \log|\Sigma_j| + n_{ij} \log|\Sigma_{ij}| + \lambda P \qquad (2)$$

where $n_{ij} = n_i + n_j$, d is the dimension of the feature vector, Σ_{ij} is the covariance matrix of the data points from two segments c_i and c_j, Σ_i is the covariance matrix of the data points from the segment c_i, Σ_j is the covariance matrix of the data points from the segment c_j, and P is:

$$P = \frac{1}{2}(d + \frac{1}{2} d(d+1)) \log(n_i + n_j) \qquad (3)$$

The ΔBIC is the distance between two Gaussian models. The negative value of ΔBIC indicates that two models fit to the data better than one common Gaussian model. The positive value of ΔBIC indicates statistical similarity of compared models. The ΔBIC is computationally effective, can compare segments with different lengths and can be used for direct audio search.

The search procedure includes some steps. The first step is a feature extraction both for query and for tested audio data. In the presented experiments mel-spectral coefficients (MFCC) plus energy and their delta MFCC are used. After the feature extraction for the query the determinant of the covariance matrix is calculated. For the search the sliding window with the fixed step is used. In each position of the window with respect to the audio signal MFFC features are extracted, ΔBIC between the analyzed segment and the query audio is calculated. The positive value of ΔBIC indicates that query audio has a good match with tested audio segment. Finally from the set of overlapped segments having positive ΔBIC with the query audio the audio segment having maximal ΔBIC is selected and their start, end and the criterion value are considered as the result of the search. The modified search algorithm in which the search window size can be varied is also used. The quality of the search depends on how correctly the tuning parameter is selected. For parameter selection the search pattern is compared with their parts. For example left part of the query audio is compared with the right part of the same audio. The tuning parameter is selected to provide minimal positive value of ΔBIC.

3. Data description

The test data is collected from several German TV broadcasters. The total duration of the data is 9 hours 52 minutes. Total jingles occurrence is 99. In Table 1 is presented information about sources of audio data, their duration and the number of occurrence of jingles in the data. For the data the sampling rate is 16kHz and 16 bits for one sample are used.

Table 1. Information about sources of audio data

	Data source	Duration (sec.)	Jingles
1	Koeln_24_04	933	6
2	Aachen_24_04	1743	4
3	Duesseldorf_24_04	1744	3
4	Bergisches_land_24_04	1745	6
5	Ruhr_24_4	1744	4
6	Dortmund_24_04	1746	4
7	Muensterland_24_04	1743	5
8	Owl_24_04	1745	4
9	Suedwestfalen_24_04	1737	4
10	Bonn_24_04	1746	5
11	Duisburg_24_04	1743	5
12	Koeln_11_04	1733	6
13	Aachen_11_04	1734	7
14	Duesseldorf_11_04	1735	4
15	Bergisches_land_11_04	1736	4
16	Ruhr_11_04	1735	4
17	Dortmund_11_04	1734	5
18	Muensterland_11_04	1733	4
19	Owl_11_04_03	958	4
20	Owl_11_04_04	1544	4
21	Suedwestfalen_11_04	1732	4
22	Bonn_11_04	1737	4

In Figure 1 some examples extracted from the tested audio data corresponding to the search patterns are presented. The patterns have variability in loudness.

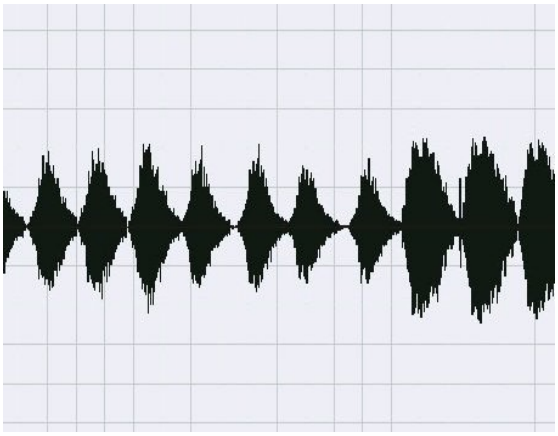

Figure 1. Some examples of concatenated search patterns

The histogram of the duration of the search pattern is presented in Figure 2.

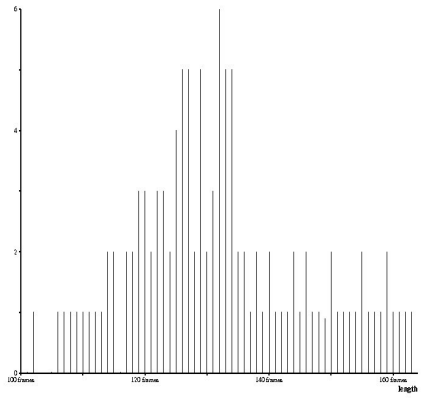

Figure 2. The histogram of the duration of the search patterns

The size of the patterns that should be found is varied from 1 to 2 seconds. The average value is equal to 1.45 seconds.

4. Direct search experiments

In this section the results of the search in 9 hours 52 minutes of the test data are presented. In the experiments the search window is varied. The total number of patterns that should be found in the data is **99**. In Table 2 the results of the search using 3 fixed size search windows are presented. The obtained hypotheses are in the column "hyp", the numbers of correctly founded search pattern are in the column "cor".

Table 2. Search query patterns using fixed size window

Features	Window	Hyp	Cor	Shift	xRT
12 mfcc+delta	1.2 sec.	130	**97**	0.25	0.023
12 mfcc+delta	1.3 sec.	124	**97**	0.25	0.023
12 mfcc+delta	1.4 sec.	120	**97**	0.25	0.024

In Table 3 the result with variable search window are presented.

Table 3. Search query patterns using variable size window

Features	Window	Hyp	Cor	Shift	xRT
12 mfcc+delta	1.2-1.4 sec	131	**98**	0.25	0.033

The algorithm with variable search window works in 1.5 times slower but does only one error in the search.

The experiments are carried out on a computer with processor Intel (R), Core™ 2 CPU 6400, 2.13 GHz.

5. Unsupervised audio events analysis

In recent years an interest has grown for environmental audio signals recognition [10], [11]. In [12] an investigation of human abilities to recognize everyday auditory scenes is presented. The human abilities are measured by the recognition accuracy, latency and by cues used to recognize events. In the experiments 34 everyday audio events such as sounds from traffic, cars, trains, street cafés, pubs, subways and etc are considered. The overall recognition rate was 66% and recognition accuracy for individual sounds varied from 0% to 100%. The overall latency of successful recognition is 20 seconds. It was shown that the most cues for recognition is described in terms of familiar sounds and events. In automatic environmental sounds recognition the performance was 58% for 24 everyday classes and 82% for 6 high level classes [11].

For audio events recognition experiments 14 environmental sounds are used. The following environment sounds are considered: airplanes, applause, car motors, car accidents, bar/restaurants, laughter, traffic, car races, town, casino, horses, weather, steps, crowds and explosions. For each sound at least 10 episodes are presented. The duration of each episode is from 5 to 10 seconds. The sounds are

obtained from [13]. In total 164 sound episodes are tested.

Each audio episode is compared with all other 164 episodes using distance based on BIC. For each pair of episodes ΔBIC is calculated. Each episode is characterized by the distance vector of the N-best non-negative largest ΔBIC. This vector describes global mutual similarity between different audio events. The vector is presented as:

$$\overline{D_i} = (\Delta BIC(i1), \Delta BIC(i2), ..., \Delta BIC(iN)) \quad (4)$$

where ΔBIC is the distances between episode i and each of the other episodes. The vector has N components according to N-best. When less than N positive ΔBIC is obtained the absent components values are equal to 0. For matrix of 164 episodes 26896 combinations are compared. The matrix is presented in Figure 3. In the picture the largest ΔBIC have diagonal elements. this is the case when episodes are compared with themselfs. This means that each episode is detected through self-comparison.

To demonstrate the similarity based on ΔBIC projection the scalar product of each par of audio events is calculated.

$$dist(\overline{D_i}, \overline{D_j}) = \sum_{k=0}^{N} \Delta BIC(i,k)\Delta BIC(j,k) \quad (5)$$

The obtained matrix is presented in Figure 4. The matrix includes much more black dots that are not diagonal but are close to diagonal. The groups of similarity are rectangular located near the diagonal. This means that similar elements are selected not only based on self detection but mostly due to mutual similarity to other elements. Numerical evaluation shows that 119 episodes from 164 (72.5%) are detected due to the similarity to the episodes of the same class. The experiments with different ΔBIC N-best using N from 1 to 10 are tested. The best results are obtained with N equal 3. Analysis of the results shows that most of the errors demonstrate the natural similarity of the analyzed episodes. For example, there are the following mismatches: car race to airplanes has (2 episodes), traffic to the airplane (1 episode), car race to the traffic (3 episodes), laughter to the crowd (1 episode), town sounds to the car race (1 episode), town sounds to the traffic (1 episode), explosion to the accident (1 episode), weather to the explosion (1 episode).

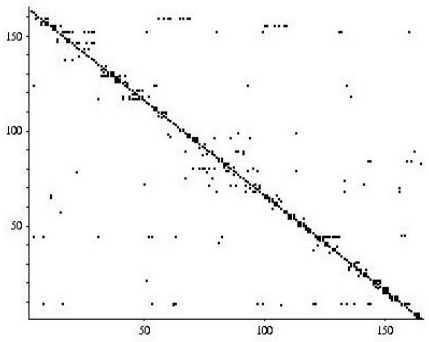

Figure 3. Comparison of audio events using ΔBIC distance

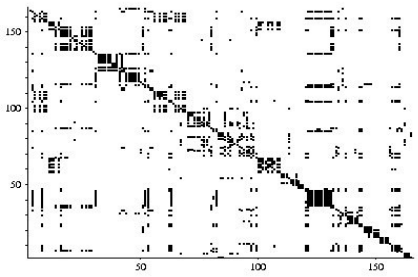

Figure 4. Comparison of audio events using ΔBIC based similarity vectors

6. Using LSI and BIC for audio data analysis

In the text retrieval and in the audio applications the technique based on LSI has been studied intensively in recent decades [14], [15], [16], [17]. LSI uses singular vector decomposition (SVD) to reduce noise in initial term-document representation and to capture hidden relationships between the terms. Using SVD the term-document matrix A is factored into the product of three matrix as:

$$A = USV \quad (6)$$

where U is an orthonormal matrix $U^TU = I$, V is also an orthonormal matrix $V^TV = I$ and S is a diagonal matrix of ordered singular values from the highest values to the lowest values. In the LSI method matrix A is approximated by using k largest singular values, factors, the other singular values are discarded.

In this paper each type of the sound using terminology of LSI is considered as a document. Each document is described by the distance of this document to other reference sounds. For distance measure ΔBIC is used. Only positive values of ΔBIC are considered. From positive ΔBIC only N-best largest are used. The ΔBIC vector corresponding to one of the document-sounds is normalized by the sum of all components of this vector. The query is an example of audio data and is a short audio segment. For each audio query the distances to the reference audio segments are calculated. Then the query Q is converted in the LSI space by:

$$\tilde{Q} = Q^T U_k S_k^{-1} \quad (7)$$

In experiments with LSI approximately 500 audio segments from 14 audio classes are used. These segments are extracted from audio source described in [13]. The distribution of the amount of tested audio episodes that include applause (1), cars (2), accidents (3), restaurants (4), laughter (5), traffic (6), car races (7), town (8), casino (9), public transportation (10), weather (11), foot steps (12), crowds (13) and explosions (14) are presented in Figure 5.

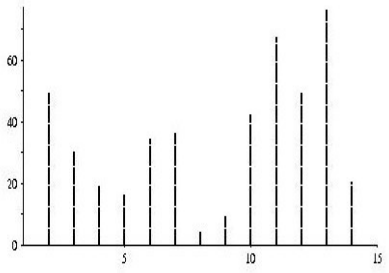

Figure 5. The distribution of the amount of different audio events used in the experiments

Each audio segment is divided into two parts. The first 2 seconds of the segment are used for training using the LSI. The rest of the segments are used for testing. These two parts are not intersected. All training segments are considered as an initial reference space. For the training 1000 seconds of audio data are used. For the testing 3000 seconds of audio data are used. The number of tested events and training segments combinations is 240100. Using BIC the term-document matrix is calculated and then is factorized using SVD. The obtained term-document matrix is approximated by using k largest singular values. The experiments with the different value of N-best ΔBIC and with the different values of the largest singular values are carried out.

In Table 4 the results of the experiments using different ΔBIC N-best are presented.

Table 4. Events recognition using different ΔBIC N-best components in the reference vector.

N-best ΔBIC	Number of factors	Correct recognition
1	350	52%
3	**350**	**76.5%**
7	350	71%

In Table 5 the results of experiments using different number of largest non-zero singular values are shown.

Table 5. Events recognition using different numbers of largest non-zero singular values

N-best ΔBIC	Number of factors	Correct recognition
3	70	52%
3	140	63.7%
3	210	70.2%%
3	250	73%
3	**320**	**76.3%**
3	**350**	**76.5%**
3	400	76%
3	450	75.9%
3	480	75.9%

The Figure 6 shows the performance of individual audio events recognition.

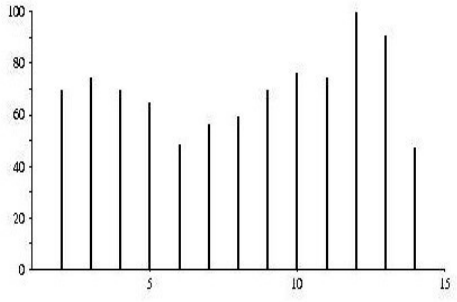

Figure 6. The results of individual audio events recognition

7. Conclusions

This paper describes applications of model selection criterion to some tasks of audio data

analysis. As an example of the model selection criterion the BIC is used. In the task of direct audio search the experiments show that model selection criterion is an effective technique which demonstrates high accuracy and high speed of the search.

In the task of unsupervised audio events analysis mutual similarity vectors based on ΔBIC are used. From 26896 possible combinations 72.5% of audio classes are detected correctly. The best results are obtained using 3-best of ΔBIC distances ignoring negative values of ΔBIC.

In the task of audio events recognition LSI technique is used. The number of training and testing data in comparison with the unsupervised audio analysis is extended. In the experiments 500 audio segments from 14 environmental audio classes are tested. The total number of combination is 240100. Each audio segment is described by reference vector that represents mutual similarity of this segment to all other reference segments. The components of this reference vector are normalized using N largest non negative values of ΔBIC. In the experiments it is found that the N-best value that provides the best performance is 3. In LSI analysis the important parameter is the number of largest non-zero singular values used for term-document matrix approximation. It is found that the best performance is obtained with 320 largest non-negative singular values.

The experiments show that the applications of model selection criterion for direct audio search, unsupervised environmental sounds analysis and sounds recognition using LSI are effective and accurate.

In the future work the audio events detection in the audio stream will be investigated. The detection in the audio stream includes segmentation and audio events separation from other unseen audio events. The considered technique will also be investigated for audio events discovery.

8. Acknowledgments

This work is done in the context of EU projects Boemie and Vitalas.

9. References

[1] J. Sen, M. Jin, D. Jang, S. Lee and C. Yoo, "Audio fingerprinting based on normalized spectral subband centroids", in *Proc. ICASSP' 2005,* 2005.

[2] W. Liang, S. Hhang and B. Xu, "A histogram algorithm for fast audio retrieval", in *Proc. ISMIR 2005,* 2005.

[3] J. Pinquier and R. Andre-Obrecht, "Jingle detection and identification in audio documents", in *Proc. ICASSP 2004,* 2004.

[4] M. Betser, P. Collen and J-B. Rault, "Audio identification using sinusoidal modeling and application to jingle detection", in *Proc. ISMIR 2007,* 2007.

[5] P. Cano, E. Batlle and J. Haitsma, "Robust identification of time scaled audio", in *Proc. of 5th IEEE Workshop on Multimedia Signal Processing*, 2002.

[6] H. Akaike, "On entropy maximization principle", *P.R. Krishnaian, editor, Application of Statistics, North-Holland, Amsterdam, Nederlands.*

[7] G. Schwarz, "Estimation the dimension of a model", *The Annals of Statistics*, 6(2), 461-464.

[8] J. Rissanen, "Stochastic complexity", *Journal of the Royal Statistical Society,* B, 49(3), 223-239.

[9] S. Chen and P. Gopalakrishman, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", in *Proc. DARPA Workshop.*

[10] S. Chu, S. Narayanan and C.-C. Jay Kuo, "Environment sounds recognition using mp-based features", in *Proc. ICASSP 2008*, 2008.

[11] A. Eronen, V. Peltonen, J. Tuomi, S. Fagerlund, T. Sorsa, G. Lorho and J. Huapaniemi, "Audio-based context recognition", *Trans. On Audio, Speech and Language Processing",* 2006.

[12] V. Peltonen, A. Eronen, M. Parviainen and A. Klapuri, "Recognition of everyday auditory scenes: potential, latency and cues", in *Proc. Of 110th Convention AES.*

[13] The sound effects library – original series", http://www.sound-ideas.com

[14] S. Deerwester, S. Dumas, T. Landauer, G. Furnas and R. Harshman, "Indexing by latent semantic indexing*", Journal of the American Society of Information Science"*, vol. 41, no. 6, pp.391-407, 1990.

[15] S.T. Dumas, "Latent semantic indexing (LSI) and TREC-2", in *Proc. Second Text Retrieval Conf (TREC-2),* 1994.

[16] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling", in *Proc. IEEE (Special Issue on Speech Recognition Understanding),* vol.88, no. 8, pp. 1279-1296, Aug. 2000.

[17]. J.R. Bellegarda, "Fast update of latent semantic space using linear transform framework", in *Proc. of ICASSP-2002*, 2002.