Query by Example Methods for Audio Signals

Marko Helén[1], *Tommi Lahti*[2]

[1]Tampere University of Technology
Institute of Signal Processing
Korkeakoulunkatu 1, FIN-33720 Tampere
Finland
Tel. +358-3-3115 3251, Fax: +358-3-3115 3857
marko.helen@tut.fi

[2]Nokia Research Center
Multimedia Technologies Laboratory
Finland
tommi.lahti@nokia.com

ABSTRACT

Various methods for query by example for audio signals are discussed in this paper. The query by example aims at automatic retrieval of audio excerpts similar to a user-provided audio sample from his/her personal audio database. Methods based on hidden Markov models, feature histograms, and likelihood ratio test are discussed.

A class based approach was adopted in defining the concept of similarity between two audio samples. For this reason two databases were constructed for the simulations. Experiments were carried out by using a high quality audio database and an audio database collected with a mobile phone. Considering the difficulty of the task the results are quite encouraging also from the future research point of view.

1. INTRODUCTION

Recently, more and more multimedia content is created and stored in digital form resulting in the growing need for accurate browsing and retrieval of text, image, audio, and video documents [1]. In the query by example approach the user provides an example of an audio signal and based on that example, similar samples are retrieved from the database. The retrieval is based on some similarity measure between the example and database samples. A difficult problem is how to define similarity itself. There is only one example signal and it is difficult even for a human to say what user means by similarity. For instance, if user gives example sample which contains male speech, it is impossible to know whether the user wants samples from the same speaker, samples from male speakers, or any speech samples.

Several related topics have been studied extensively. In content-based classification there are predefined classes for instance, for speech, music, and environmental sounds. Test signal is then classified into one of the classes. Common underlying techniques include for example neural networks [2] and HMMs [3].

Another related topic is event matching. The purpose is to find certain audio events from an audio stream or from long samples of audio. Applications are for instance finding a certain scenes, like car-chasing or gunplay, from a movie [4] or extracting highlights like laughter, cheer, and applause [5]. HMMs or support vector machines (SVM) are typically used for classification.

Third related subject is auditory scene recognition, which tries to detect the acoustic environment at certain time. Usually this means that first the segment boundaries, where the change from one environment to another happens, have to be identified. Secondly the identified segments must be classified into some acoustic environment. Tuomi applied Gaussian mixture models(GMMs) and HMMs for this task [6].

The paper is organized as follows. Section 2 describes the system overview. In Section 3, the methods for query by example are proposed. Section 4 deals with simulation experiments. Finally, Section 5 gives conclusions.

2. SYSTEM OVERVIEW

The problem of query by example can be roughly separated into two sub-problems. The first is to find such features that reveal the similarities or differences between samples. Second sub-problem is, how to measure the distance between the feature vectors.

A block diagram of the system is presented in Fig. 1. First, the feature vectors are estimated for both, the example signal from the user, and for the database signal. One by one each database signal is compared to the example signal. If similarity criterion is fulfilled, the database sample is retrieved. Methods for comparing two samples are described in the next section.

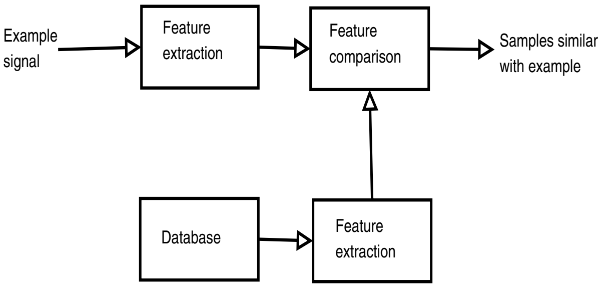

Fig. 1. A block diagram of the overall system.

2.1 Feature Extraction

Signal is divided into frames and feature vector is calculated for each frame. Both, features which describe the frequency content of the signals, and features which describe the temporal characteristics were used in the system. The feature set includes: MFCCs, zero crossing rate, crest factor, spectral centroid, noise likeness, pause rate, power, power variance, harmonic ratio, lag, spectral spread, and spectral flux. Features are normalized over the whole database with zero mean and unity variance. Several different types of features for audio signals are described by Peeters [7].

Linear Discriminant Analysis (LDA) was also used to reduce the correlation between the feature components in the feature set. First, the original feature set is calculated. Then, LDA finds a projection of features to such a feature space which maximizes the ratio of ”between” and “within” class covariance.

3. METHODS FOR SIMILARITY COMPARISON

Since there are not many studies about query by example, several methods were tested in this paper. Here are the principles of those methods.

3.1 HMM

The traditional HMMs have been successfully used for various classification tasks [3][4][5][6]. Here, they were tested also for query by example task. Two models are generated, one for the example signal and one for the background. The model for the example is trained using only the feature vectors from the example sample. Model for background is trained using the data from the whole database. Similar approach was applied in audio segment retrieval by Velivelli et al. [8].

During classification, the likelihoods for each database sample belonging to these two models are estimated. Samples which have higher likelihood for the example model are assumed to be similar with the example and the samples which have higher likelihood for background model are decided to be not similar.

The main problem here is that there is only one example signal available for training. This is why we tried to increase the number of training samples artificially. The artificial training samples were generated by adding the random Gaussian noise to the feature vectors of the example signal. Then HMM was trained using these generated example signals. However, the results for normal HMM were better than results using the noise adding.

3.2 Likelihood Ratio Test

In the likelihood ratio test (LRT), observations for the example and database signal are combined together. Then n-component and $2n$-component GMMs are calculated from the feature vectors of these longer samples. Likelihoods for both models are calculated and if the score for the n-component GMM model is higher than for the $2n$-component GMM, the original samples are assumed to be similar. A block diagram of the system is provided in Fig. 2.

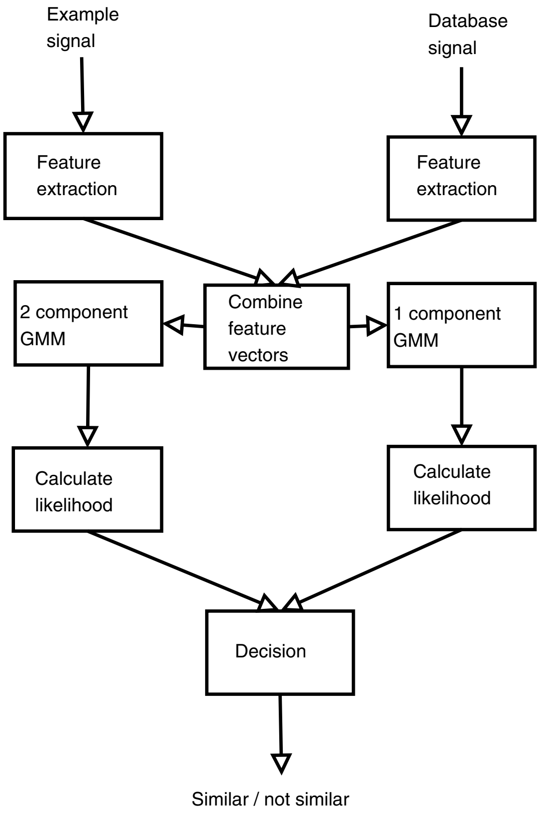

Fig. 2. A block diagram for the likelihood ratio test.

The number n for the components is chosen using Minimum Description Length (MDL) and expectation-maximization algorithm [9]. MDL optimizes the number of GMM components for model given the samples. For the paper also fixed number of components was tested. Both approaches gave nearly same results.

3.3 Histogram Model

The feature histogram is generated from the frame based feature vectors. First, based on the features calculated from all the samples in the database, the centers of each quantization level is estimated using Linde-Buzo-Gray (LBG) vector quantization algorithm [10]. Then example signal and database signal histograms are generated by assigning feature vectors from each frame to the closest quantization level. As a result one feature histogram from each sample is generated. The distance between histograms is calculated and if the distance is below a threshold, the samples are assumed to be similar [11]. A diagram of the histogram model is presented in Fig.3.

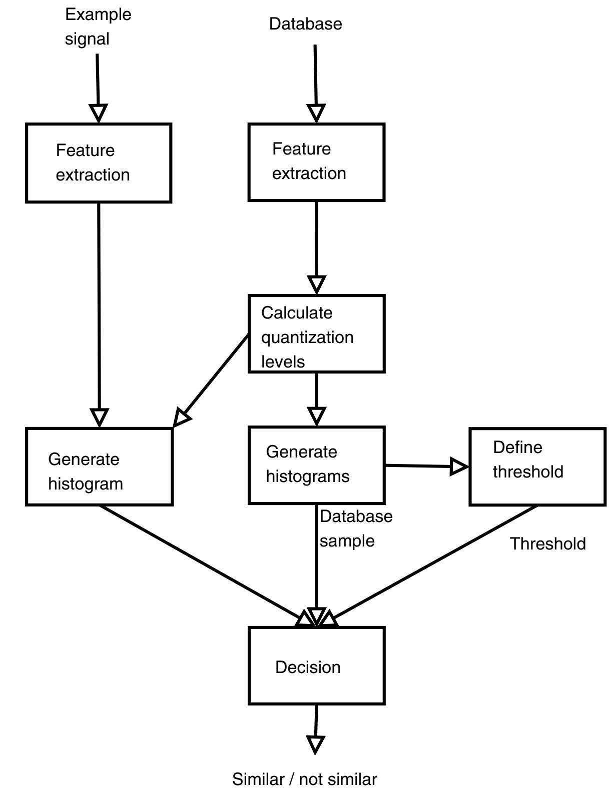

Fig. 3. A block diagram of the histogram method.

In this system the amount of acceptable fluctuation between the ”similar” samples can easily be adjusted with the threshold. On the other hand, this method does not consider the changes in time like HMMs do. The threshold is calculated as follows:

$$threshold = mean(d) + c \cdot std(d) , \qquad (1)$$

where d contains all the distances between the samples in the database and c is an empirically determined constant[12].

Different measures for histogram distances were tested. These included: L1, L2, and Linf –norms, and Kullback-Leibler distance.

4. SIMULATION EXPERIMENTS

The performance of the proposed methods was measured through the following simulations. Tests were carried out using both mobile audio database and the 16 kHz high quality (HQ) database. The mobile audio database contained 1074 audio signals. These were divided into 2751 three second samples, which were used in the simulations. The signals were manually annotated into 16 predetermined classes. Samples falling into each class were considered to be similar. The classes and the number of samples in each class are listed in Table 1. In the HQ audio database there was data from 4 classes, 60 samples from each: music, speech, environmental noise, and constant noise.

4.1 Evaluation Procedure

One sample at the time is drawn from the database serving as an example sample and this is compared against all the other samples in database. The same procedure is repeated for all the samples. This way there are $n(n$-1) comparisons, where n is the number of samples in the database. If in the comparison the two samples are considered to be similar and they are labeled in the same class, the database sample is seen as correctly retrieved from the database.

The results are presented here as an average value of recall and precision rates. Recall means how large portion of the similar samples was found from the database:

$$recall(class) = \frac{N_{ccs}}{n_{class}(n_{class}-1)} , \qquad (2)$$

where n_{class} is the number of samples belonging to the class, and N_{ccs} means the number of correctly classified samples from this class.

Precision gives the portion of correctly classified samples into a certain class over all possible samples classified in the database into the same class:

$$precision(class) = \frac{N_{ccs}}{N_D} , \qquad (3)$$

where N_D means the total number of samples in the database that are classified into the class when the example sample(s) are drawn from that particular class.

Unfortunately, there are different number of samples in different classes in mobile audio database and this affects the precision rate in a such way that classes which have more samples have better precision simply because there is not so many wrong samples than for the smaller classes. This, however, does not affect the recall rates.

In histogram model 8 quantization levels were used, threshold variable was set to 1.2 and L1-norm was selected as a distance measure. The HMM models had 5 states each state composed of 2 densities.

4.2 Results

The results for HMM, histogram and LRT methods using mobile audio database are presented in Table 1. Results from HQ database are presented in Table 2. The latter ones are significantly better than with mobile audio database. The reason is that there is a clear difference in signal quality between the two databases. Also, HQ database has fewer classes which has an effect on precision rate as there are simply less false samples when performing a query.

Comparing the methods and interpreting the results shown in the tables below is not trivial. Both recall and precision rates need to be considered at the same time when doing the judgement. For example, the LRT method seems to outperform the other methods in terms of recall rates on mobile audio data. Examining the precision rates closer, however, there seem to be indications that LTR method badly over-classifies almost all the samples to be similar with the example sample. Also, on the HQ data both the recall and precision rates tend to be worse in most of the cases compared to the other methods.

It is not evident which of the methods, HMM or histogram method would eventually perform better in the query by example task. An advantage with the histogram model is that the threshold can easily be adjusted. If higher precision is required the threshold is raised and for better recall the threshold can be lowered.

Variations in acoustic expressions between the samples inside the individual classes are quite different between the classes. In some classes for instance, male speech, the variation is not very high but on the other hand, animal sounds contains samples which vary quite a lot. As a consequence, there is a great variation in precision and recall rates between different classes.

5. CONCLUSIONS

In this paper, several query by example methods were introduced and tested for audio signals. The histogram and HMM methods performed the best. In some cases histogram method is probably more convenient, because the similarity threshold can be easily adjusted. The best results so far were: recall 25%, precision 28% from the HMM method and recall 38%, precision 18% from the feature histogram method on mobile data. For HQ data the recall and precision rates for the HMM was 27% and 83% respectively and for histogram method 52% and 68% respectively.

Table 1. Results from different methods with mobile audio database. Results are presented in terms of recall/precision percents.

Class(Number of samples)	HMM	Histogram	LRT
male speech (560)	16/54	36/50	77/11
Female speech (257)	24/32	39/28	49/8
laughter (13)	15/4	31/1	40/2
singing (16)	43/3	67/3	28/2
whistling (84)	40/71	51/20	50/8
pop music (1054)	10/80	13/51	74/11
classical music (168)	19/13	21/12	41/7
car (131)	45/43	50/33	62/9
airplane (24)	13/5	15/15	67/5
train (110)	36/39	46/23	85/11
motorbike (22)	16/4	19/2	46/3
City noise (48)	30/21	63/11	55/9
Animal sounds (190)	11/18	13/7	36/8
applause/cheer (43)	48/24	84/17	61/8
Blast sounds (30)	8/3	15/2	49/3
average	**25/28**	**38/18**	**55/7**

Table 2. Results from different methods with HQ database. Results are presented in terms of recall/precision percents.

Class	HMM	Histogram	LRT
music	36/92	36/70	4/4
environmental noise	23/40	66/58	55/24
constant noise	27/100	71/60	46/18
speech	21/100	35/85	57/24
average	**27/83**	**52/68**	**41/18**

REFERENCES

[1] S. Kiranyaz, *Advanced Techniques for Content-Based Management of Multimedia Databases*, Ph.D. thesis, Tampere University of Technology, Finland, 2005.

[2] X. Shao, C. Xu and M.S. Kankanhalli, "Applying neural networks on the content-based Audio Classification," in *Proc. Fourth Int. Conference on Information, Communicatoins and Signal Processing*, Singapore, Dec., 2003, vol. 3, pp. 1821-1825.

[3] T. Zhang, and C.-C.J. Kuo, "Hierarchical Classification of Audio Data for Archiving and Retrieving," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'99),* Phoenix, USA, March, 1999, vol. 6, pp. 3001-3004.

[4] W.-T. Chu, W.-H. Cheng, and J.-L. Wu, "Generative and discriminative modeling toward semantic context detection in audio tracks," in *Proc. 2005 Int. Multi-Media Modeling Conference (MMM 2005)*,Melbourne, Australia, January, 2005.

[5] R. Cai, and L. Lu, Hong-Jiang Zhang, Lian-Hong Cai, "Highlight sound effects detection in audio stream," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME03),* USA, 2003, vol. 3, pp. 37-40.

[6] J. Tuomi, *Audio-Based Context Tracking*, M.Sc. thesis, Tampere University of Technology, 2004.

[7] G. Peeters, *A Large Set of Audio features for sound description (similarity and classification) in CUIDADO project*, CUIDADO I.S.T Project Report 2004, 2004.

[8] A. Velivelli, C. Zhai, and T. S. Huang, "Audio segment retrieval using a short duration example query," *in Proc. ICME 04*, Taipei, Taiwan, June, 2004, vol. 3, pp. 1603-1606.

[9] M. Figueiredo, J. Leito, and A. K. Jain, "On fitting mixture models," in E. Hancock, and M. Pellilo (Editors), *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 54 - 69, Lecture Notes in Computer Science, vol. 1654, Springer Verlag, 1999.

[10] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," in *IEEE Transactions on Communications*, vol. COM-28, pp. 84-95, Jan. 1980.

[11] G. Smith, H. Murase, and K. Kashino, "Quick audio retrieval using active search," in *Proc. ICASSP'98*, Seattle, Washington, USA, May, 1998, vol. 6, pp. 3777-3780.

[12] K. Kashino, T. Kurozumi, and H. Murase, "A quick search method for audio and video signals based on histogram pruning," in *IEEE Transactions on Multimedia*, vol. 5, no. 3, Sep. 2003.