

QUERY BY EXAMPLE OF AUDIO SIGNALS USING EUCLIDEAN DISTANCE BETWEEN GAUSSIAN MIXTURE MODELS

Marko Helén, Tuomas Virtanen

Tampere University of Technology
Institute of Signal Processing
Korkeakoulunkatu 1, FIN-33720 Tampere, Finland

ABSTRACT

Query by example of multimedia signals aims at automatic retrieval of media samples from a database, which are similar to a user-provided example. This paper proposes a method for query by example of audio signals. The method calculates a set of acoustic features from the signals and models their probability density functions (pdfs) using Gaussian mixture models. The method measures the similarity between two samples using the Euclidean distance between their pdfs. A novel method for calculating the closed form solution of the distance is proposed. Simulation experiments show that proposed method enables higher retrieval accuracy than the reference methods.

Index Terms— Acoustic signal processing, database query processing, feature extraction.

1. INTRODUCTION

The management of ever growing multimedia databases is time consuming when done completely manually. This is why automatic systems are required to lighten the job. Query by example aims at automatical retrieving of samples from a database, which are similar to the example provided by the user. For example, the user gives a sample of a dog barking and system returns all the samples from the database which contain dog barking. Query by example differs from supervised classification in the sense that there are no predefined classes for which the system could be trained. Therefore, unless the similarity is defined beforehand according to a certain criterion, query using only one example is not a well defined problem. Consider the situation where the user gives an example of male speech: without predefining the similarity metric, it is impossible to know whether the user wants samples from the same speaker, from all male speakers, or speech in general.

The existing methods usually overcome the above limitation by representing the samples using a set of features which have been found to correlate with the perceptual similarity. Thus, query by example is usually done in the following way [1, 2, 3]: first, features are extracted from the example and all the samples in the database. Second, the distances between the feature vectors of the example and the database samples are estimated using a certain distance metric. Finally, database samples having the shortest distance to the example are retrieved.

Spevak and Favreau [1] used the average Euclidean distance between the features. They also applied self-organizing maps (SOM) to project the high-dimensional feature data into a two-dimensional space, and calculated the Euclidean distance between the SOMs. Gabbouj et al. [4] used a method, where samples were first clas-

sified into four main categories and then searched for similar samples only within the samples of the same main category. Helén and Lahti [2] used following three different methods. In the feature histogram method the feature vectors were quantized and the similarity was measured by the distance between their histograms. The hidden Markov model (HMM) method generates a HMM for the example and then uses the likelihood of the database sample to estimate whether it is more likely to be generated by the example model than by a background model. The likelihood ratio test models the combination of the example and a database sample using n -component and $2n$ -component GMMs and retrieves samples for which the n -component likelihood is higher.

In this paper we estimate the similarity of two samples by measuring the difference between their probability density functions (pdfs) p_1 and p_2 of the features. However, the features are continuous-valued which makes the estimation of their pdfs difficult. Previously this has been solved by quantizing the observation values and calculating their counts within each quantization level to obtain observation histograms [5]. The drawback in the quantization is that if two observations fall into different quantization levels, they are regarded as different even when they are closely spaced. To overcome this limitation, we model the continuous pdfs of the samples by Gaussian mixture models. The existing methods operating on continuous pdfs estimate the similarity by the likelihood [6] that the database sample is generated by the pdf of the example. However, the likelihood is not ideal in the sense that it is negative even for identical samples, whereas the distance between the pdfs can reach zero. We propose a method for calculating the closed-form solution for the Euclidean distance between two GMMs. The proposed similarity measure is shown to produce higher accuracy in the simulations than the existing methods.

The paper is organized as follows. Section 2 describes the feature extraction and modeling, Section 3 proposes the similarity measure. Section 4 introduces different ways to retrieve the samples. Section 5 gives experimental results and comparisons to the other methods.

1.1. System overview

An overview of the system is illustrated in Fig. 1. First, the example signal given by the user is divided into frames and a set of features is extracted within each frame. Second, a GMM which models the feature distribution is estimated using the expectation maximization (EM) algorithm. The same set of features and a GMM is estimated for each sample in the query database beforehand. Third, the example signal is compared against the database signals one at the time and similarity between all pairs is estimated by the Euclidean distance between their pdfs. Finally, when all the similarity values are

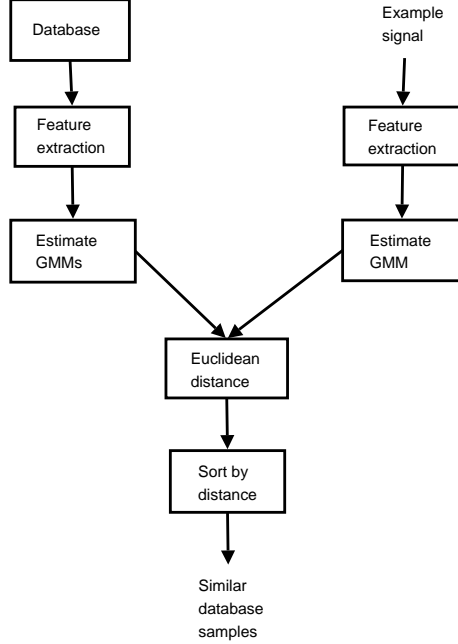


Fig. 1. Overview of the query by example system.

calculated a decision is made regarding the similarity of the samples to the example and those considered similar are returned to the user.

2. FEATURE EXTRACTION AND MODELING

The goal of feature extraction is to describe the perceptual properties of a signal using a small number of parameters. The input signal is divided into 46 ms frames and a set of features is extracted in each frame. The frequency content of the frame is described using three Mel-frequency cepstral coefficients, spectral centroid, noise likeness [7], spectral spread, spectral flux, harmonic ratio [8], and maximum autocorrelation lag. Temporal characteristics of the signal are described using zero crossing rate, crest factor, total energy, and variance of instantaneous power. Each feature is normalized to have zero mean and unity variance over the whole database. The total number of features is $N = 13$, and \mathbf{x} is used to denote the feature vector of length N within each frame.

2.1. Gaussian Mixture Model for the Features

The distribution $p(\mathbf{x})$ of the features of each sample is modeled using a Gaussian mixture model (GMM), defined as

$$p(\mathbf{x}) = \sum_{i=1}^I w_i \mathcal{N}_i(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (1)$$

where w_i is the weight of the i^{th} component, I is the number of components, and \mathcal{N}_i is the multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$. The weights are non-negative and sum to unity.

Two methods for estimating the parameters of the GMMs were tested. The first uses the expectation maximization (EM) algorithm to estimate the means and variances for a fixed number of components. It should be noted that the variances have to be restricted

above a relatively high fixed minimum level, since low-variance components would dominate the measure. The second method uses the Parzen-window [9, pp. 164-174] approach which assigns a GMM component with fixed variance for each observation so that I equals the number of frames, $\boldsymbol{\mu}_i$ is the feature vector within frame i , $\sigma_{i,n}$ is fixed, and $w_i = 1/I$.

3. PROPOSED SIMILARITY MEASURE

The similarity of two samples is measured by the square of the Euclidean distance e between their distributions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$. This is obtained by integrating the squared difference over the whole feature space:

$$e = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} [p_1(\mathbf{x}) - p_2(\mathbf{x})]^2 dx_1 \dots dx_N \quad (2)$$

To simplify the notation, we rewrite the above multiple integral as

$$e = \int_{-\infty}^{\infty} [p_1(\mathbf{x}) - p_2(\mathbf{x})]^2 d\mathbf{x} \quad (3)$$

in the following.

3.1. Closed-Form Solution for the Euclidean Distance Between GMMs

The Euclidean distance (3) can be written as $e = e_{11} + e_{22} - 2e_{12}$, where the three terms are defined as

$$e_{11} = \int_{-\infty}^{\infty} [p_1(\mathbf{x})]^2 d\mathbf{x}, \quad (4)$$

$$e_{22} = \int_{-\infty}^{\infty} [p_2(\mathbf{x})]^2 d\mathbf{x}, \quad (5)$$

and

$$e_{12} = \int_{-\infty}^{\infty} p_1(\mathbf{x})p_2(\mathbf{x}) d\mathbf{x}. \quad (6)$$

All the above terms are definite integrals of the product of two GMMs, for which the closed-form solution can be obtained. First, let us write the product of two normal distributions \mathcal{N}_1 and \mathcal{N}_2 as

$$\begin{aligned} \mathcal{N}_1(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}_2(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \\ \frac{1}{(2\pi)^N \sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \exp \left[-\sum_{n=1}^N \left(\frac{(x_n - \mu_{1,n})^2}{2\sigma_{1,n}^2} + \frac{(x_n - \mu_{2,n})^2}{2\sigma_{2,n}^2} \right) \right] \end{aligned} \quad (7)$$

where $\mu_{k,n}$ is the n^{th} entry of mean vector $\boldsymbol{\mu}_k$, $k \in \{1, 2\}$, and $\sigma_{k,n}^2$ is its variance.

Second, we use the identity

$$\begin{aligned} \int_{-\infty}^{\infty} \exp \left[-\frac{(a-b)^2}{e^2} - \frac{(a-c)^2}{f^2} + g \right] da \\ = \frac{\sqrt{\pi}|e|f}{\sqrt{e^2+f^2}} \exp \left[-\frac{(c-b)^2}{e^2+f^2} + g \right] \end{aligned} \quad (8)$$

and integrate (7) N times, with respect to all the entries of \mathbf{x} to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \mathcal{N}_1(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}_2(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) d\mathbf{x} \\ = \frac{1}{(2\pi)^{N/2} \prod_{n=1}^N \sqrt{\sigma_{1,n}^2 + \sigma_{2,n}^2}} \exp \left[-\frac{1}{2} \sum_{n=1}^N \frac{(\mu_{1,n} - \mu_{2,n})^2}{\sigma_{1,n}^2 + \sigma_{2,n}^2} \right] \end{aligned} \quad (9)$$

Let us denote the integral (9) of the product of the i^{th} component of GMM $k \in \{1, 2\}$ and the j^{th} component of GMM $m \in \{1, 2\}$ by $Q_{i,j,k,m}$.

The values for the terms e_{11} , e_{22} , and e_{12} in (4)-(6) can now be calculated as

$$e_{11} = \sum_{i=1}^I \sum_{j=1}^I w_i w_j Q_{i,j,1,1}, \quad (10)$$

$$e_{22} = \sum_{i=1}^J \sum_{j=1}^J v_i v_j Q_{i,j,2,2}, \quad (11)$$

and

$$e_{12} = \sum_{i=1}^I \sum_{j=1}^J w_i v_j Q_{i,j,1,2}, \quad (12)$$

where w_i and w_j are the weights of the i^{th} and j^{th} component of GMM 1, v_i and v_j are the weights of the i^{th} and j^{th} component of GMM 2, and I and J are the number of components in GMM 1 and GMM 2 respectively.

The Euclidean distance $e = e_{11} + e_{22} - 2e_{12}$ is used as a distance measure for the similarity: the smaller the distance, the more similar are the samples.

4. QUERY OUTPUT

When the similarity estimates are received, there are two application-dependent main possibilities how to return the results to the user. The first one is to sort the signals in order of similarity and retrieve a fixed number of most similar samples to the user. The drawbacks are that there is a possibility that some of the received samples are very different from the example, since the fixed number of samples is retrieved. Furthermore, the whole database has to be queried before the results can be presented.

The other possibility is to retrieve all the samples having the distance below a predefined fixed threshold, which may be defined manually by the user or calculated automatically using the distances of the database samples. This enables returning similar samples during the query processing. The disadvantage of this method is that adjusting the threshold may not be straightforward and it might require user feedback.

5. SIMULATION EXPERIMENTS

The performance of the proposed similarity measure was tested against existing methods. The feature histogram method uses vector quantization to quantize feature vectors, generates feature histograms, and estimates the Euclidean distance between them [2]. The GMM method uses either the EM algorithm or Parzen window method to estimate a GMM for the example and evaluates the likelihood of the database sample. The methods are called *histogram*, *EM-likelihood*, and *Parzen-likelihood* in the following. The proposed Euclidean distance for the GMMs estimated using the EM algorithm and Parzen window method are called *EM-Euclidean* and *Parzen-Euclidean*. The number of Gaussians used in the EM algorithm was 8. In the Parzen-Euclidean method different variances were tested and $\sigma^2 = 2$ which produced approximately the best results was used in the final simulations (using a fixed value is possible because the variances of the features have been normalized to unity). In the EM-Euclidean method the feature variances were restricted above unity.

Class	Parzen-Eucl.	Histogram	EM-likel.
stationary noise	78 / 79	53 / 24	70 / 42
music	58 / 58	45 / 52	53 / 76
environmental noise	58 / 44	63 / 99	48 / 59
speech	86 / 100	67 / 49	90 / 84
average	70 / 70	57 / 57	65 / 65

Table 1. Precision / recall for different classes.

Simulations were carried out using a database consisting of 240 samples with 16 kHz sampling rate. The lengths of the samples varied between 5 and 30 seconds. The samples were manually annotated into 4 classes: speech, music, environmental noise, and stationary noise.

5.1. Evaluation procedure

One sample at a time was drawn from the database to serve as an example for a query and the rest were considered as the database. A database sample was considered correctly retrieved, when it was annotated into the same class as the example. The query was repeated using each of the S samples as the example, resulting in altogether $S(S-1)$ pairwise comparisons. The number of correctly retrieved samples c_u was calculated for each class $u \in \{1, 2, 3, 4\}$. The ratio of correctly retrieved samples to all the comparisons is given by the average of recall of each query

$$\text{recall}(u) = \frac{c_u}{S_u(S_u - 1)}, \quad (13)$$

where S_u is the total number of samples in the class u . The ratio of correctly classified samples to all the samples r_u retrieved for class u examples is given by the precision

$$\text{precision}(u) = \frac{c_u}{r_u}. \quad (14)$$

The overall precision and recall were estimated for the whole database as the average of the class-wise precision and recall.

5.2. Results

Overall recall and precision for the tested methods with different values of the threshold are illustrated in Figure 2. The Parzen-Euclidean method produces the best results on the average. Furthermore, both methods which use the Euclidean distance produce better results than the other methods.

Table 1 presents the results for three methods, when the database samples having the distance below a fixed threshold were considered similar. The thresholds were set to the level which produces equal average precision and recall values. The proposed Parzen-Euclidean method produces 5 percent units higher average precision and recall than the EM-likelihood method and 13 percent units higher average than the histogram method. On average the speech class samples are retrieved more accurately than the others.

Fig. 3 presents the results for the proposed method and different classes, when n most similar samples to the example are considered. The number of samples in each class was 60 and therefore values between 1 and 60 were used for n . It can be seen in the figure that the precision is very high when less than 10 most similar samples are retrieved.

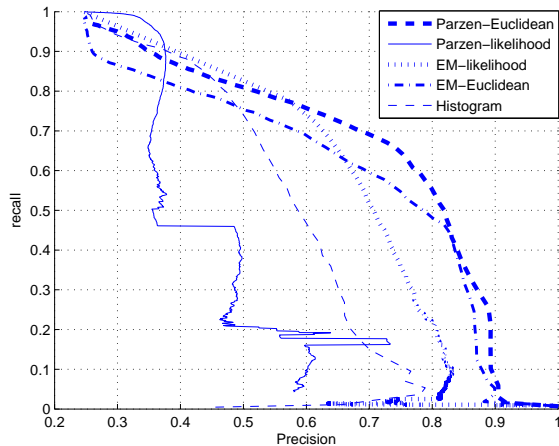


Fig. 2. The average precision and recall for the tested algorithms at different threshold values.

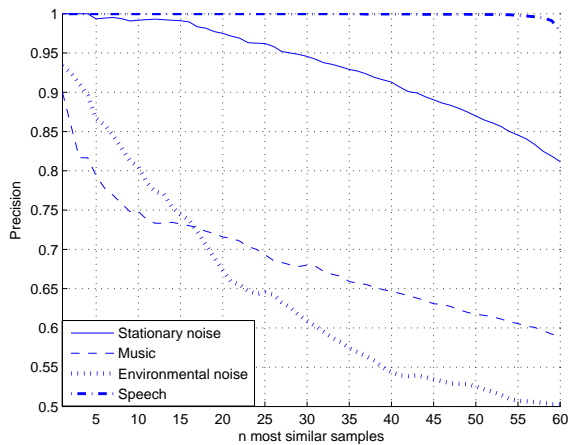


Fig. 3. Results of the Parzen-Euclidean method when a fixed number of most similar samples were retrieved.

6. CONCLUSIONS

In this paper, a novel approach to the query by example of audio signals was presented. We model the continuous pdfs of the acoustic features using GMMs, which enables calculating the pdfs without quantizing the features. A novel method for calculating the Euclidian distance between the pdfs of two GMMs is proposed and the measure was successfully used as a similarity measure in the described application.

The proposed method was tested against the previous query by example methods based on histograms of quantized features and likelihoods of GMMs. The proposed method enabled higher precision and recall rates than the reference methods. In the comparison to the likelihoods of GMMs the average precision and recall rates were increased from 65 to 70.

The basic problem in query by example using only a single example is the definition of similarity itself. Based on only one example it is difficult even for a human to say what the user means

with similarity. Therefore, the future work will consider taking the feedback from a user. When the first query is done, the user could guide the algorithm by telling which retrieved samples were correct or which were not and the system could learn from this feedback. This way the system gains information regarding the users idea of similarity. A new query could then be done based on this improved knowledge.

7. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program 2006 - 2011).

8. REFERENCES

- [1] C. Spevak and E. Favreau, "Soundspotter - A Prototype System for Content-Based Audio Retrieval," in *Proc. 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sept. 2002.
- [2] M. Helén and T. Lahti, "Query by Example Methods for Audio Signals," in *Proc. 7th IEEE Nordic Signal Processing Symposium*, Reykjavik, Iceland, June 2006, pp. 302–305.
- [3] S. Ravela and R. Manmatha, "Retrieving Images by Similarity of Visual Appearance," in *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, San Juan, Puerto Rico, June 1997, pp. 67–74.
- [4] M. Gabbouj, S. Kiranyaz, K. Caglar, E. Guldogan, O. Guldogan, and F. Ahmad, "Audio-Based Multimedia Indexing and Retrieval Scheme in Muvis Framework," in *Proc. Symposium On Intelligent Signal Processing and Communication Systems (IS-PACS)*, Awaji Island, Japan, 2003.
- [5] K. Kashino, T. Kurozumi, and H. Murase, "A Quick Search Method for Audio and Video Signals Based on Histogram Pruning," *IEEE Transactions on Multimedia*, vol. 5, no. 3, Sept. 2003.
- [6] E. Pampalk, "Computational Models of Music Similarity and their Applications in Music Information Retrieval," Ph.D. dissertation, Technische Universitat, Wien, 2006.
- [7] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of Drum Tracks From Polyphonic Music Using Independent Subspace Analysis," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003.
- [8] J. J. Burred and A. Lerch, "A Hierarchical Approach to Automatic Musical Genre Classification," in *Proc. 6th International Conference on Digital Audio Effects (DAFX)*, London, UK, Sept. 2003.
- [9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, Inc., 2001.