PROBABILISTIC MODEL BASED SIMILARITY MEASURES FOR AUDIO QUERY-BY-EXAMPLE

Tuomas Virtanen and Marko Helén

Tampere University of Technology
tuomas.virtanen@tut.fi, marko.helen@tut.fi

ABSTRACT

This paper proposes measures for estimating the similarity of two audio signals, the objective being in query-by-example. Both signals are first represented using a set of features calculated in short intervals, and then probabilistic models are estimated for the feature distributions. Gaussian mixture models and hidden Markov models are tested in this study. The similarity of the signals is measured by the congruence between the feature distributions or by a cross-likelihood ratio test. We calculate the Kullback-Leibler divergence between the distributions by sampling the distributions at the points of the observations vectors. The cross-likelihood ratio test is evaluated using the likelihood of the first signal being generated by the model of the second signal, and vice versa. Simulations were conducted to test the accuracy of the proposed methods on query-by-example of audio. On a database consisting of of speech, music, and environmental sounds the proposed methods enable better retrieval accuracy than the existing methods.

1. INTRODUCTION

Measuring the similarity of two audio signals has several applications in audio database management tasks. For example, it enables clustering the data into conceptually homogeneous clusters in an unsupervised manner [1, 2] and making database queries by an user-provided example [3]. It also forms a basis for supervised support vector machine classifiers [4, 5]. Similarity measures can also be used to segment and cluster segments of individual speakers from signals consisting of several speakers [6].

Whereas several efficient perceptual distortion measures between individual stationary spectra have been proposed, measuring the similarity of two audio signals consisting of multiple frames is a significantly more challenging task. There are several “layers” in a signal in which a human can focus into: for example a human can judge the similarity of male speech signal by the topic of the speech, by the speaker identity, or by any sounds on the background. Therefore, it is unlikely that a single measure could be used to measure the similarity of long audio signals.

The most commonly used acoustic features for audio classification are calculated in short frames. Previous studies have shown that the similarity of two long audio signals can be efficiently measured by the difference of their feature distributions – the smaller the difference, the more similar are signals. For example, Mandel and Ellis [5] calculated the average feature vector of each signal and measured their similarity by the Euclidean distance between the average feature vectors. Many systems use a parametric model such as Gaussian mixture model for the distributions [4,7] and then measure difference between them.

A commonly used similarity measure in speaker segmentation and clustering [1, 6] is the likelihood ratio test. It calculates the likelihood that the signals are generated by the same model and the likelihood that the signals are generated by individual models, and then measures their similarity by the ratio of the likelihoods.

In this paper we apply distance measures between feature distributions and likelihood ratio tests to query-by-example of audio signals. We propose an accurate measure to approximate the symmetric Kullback-Leibler divergence between two distributions, and show its relationship to the likelihood ratio test. We extend the likelihood ratio test to hidden Markov models to facilitate the use of temporal information about the signals. Simulation experiments show that the proposed methods enable better accuracy in audio query-by-example than the existing methods.

The paper is organized as follows: Section 2 presents Gaussian mixture model and hidden Markov model used for the feature distributions, Section 3 presents the measures between feature distributions, and Section 4 the cross likelihood ratio test. Section 5 presents the simulations.

2. MODELS FOR ACOUSTIC FEATURES

In audio classification, the most commonly used acoustic features are calculated in short (about 40 ms) frames, and they typically characterize the shape of the short-time spectrum. Temporal information about the signal is commonly modeled using temporal derivatives of the features or by specific models such as the hidden Markov model. There are also features which carry information about the whole signal such as the pause rate, but usually these can be derived from statistical properties of frame-wise features.

At this point we do not commit ourselves to a certain set of features, since the methods can be applied to any features which are calculated in short intervals. Let us denote the set of features calculated in frame t by vector $\mathbf{x}_t$, and the feature vector sequence of a whole signal as matrix $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_T\right]$, T being the number of frames. The distributions of the feature vectors are modeled by Gaussian mixture models and their temporal evolution by hidden Markov models. If a model is not too complex, we can estimate its parameters from an individual audio signal.

The Gaussian mixture model (GMM) models the probability distribution function of a feature vector as a weighted sum of multivariate normal distributions. Here λ denotes the whole parameter set of a particular GMM, and $p(\mathbf{x}|\lambda)$ denotes the distribution parameterized by the GMM. The distribution of a whole signal is typically obtained by assuming that the frames are statistically independent from each other, so that the frame-wise distributions

This work was supported by the Academy of Finland, project No. 5213462 (Finnish centre of Excellence program 2006-2011).

can be multiplied. Also other methods for combining the framewise probabilities have been used, but in our studies the statistical independent assumption produced the best results. Thus, the distribution of a whole signal $\mathbf{X}$ can be written as

$$p(\mathbf{X};\lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t|\lambda). \tag{1}$$

By assuming that the individual frames of a signal are independent observations, we can train a GMM for a signal by fitting the parameters to the observations using the expectation maximization algorithm.

A hidden Markov model (HMM) consists of a set of states, each of which has an individual probability distribution for producing an observation. It enables modeling the temporal evolution of signals by using a hidden state variable, which evolves over time by changing from state to another. Here we use HMMs as an extension to GMMs, and model the state emission probabilities by GMMs. Provided that the number of states times the number of Gaussians is significantly less than the number of frames, we can train a HMM from one signal. For simplicity, we denote the parameter set of a particular HMM also by λ.

The likelihood of an observation can be estimated by the expectation value over different state transition paths. In this study we obtained better results by using the most likely state transition path estimated by the Viterbi algorithm.

3. DISTRIBUTION DISSIMILARITY MEASURES

Measuring the difference between the feature distributions of two signals has turned out to be an efficient way for measuring their similarity. The earlier studies vector quantized the continuous-valued feature vectors [8], and calculated the feature histograms, after which any difference measure between the histograms could be used. Quantizing feature vectors is a source of inaccuracy and therefore several studies [4, 5, 7] have recently modeled continuous distributions using GMMs and measured the similarity by their congruence.

Given feature sequence matrices $\mathbf{A}$ and $\mathbf{B}$ of two signals, we train their GMMs λ_a and λ_b, and denote the probability distributions as $p(\mathbf{x}|\lambda_a)$ and $p(\mathbf{x}|\lambda_b)$, respectively. In audio classification, difference between the distributions has previously been measured by the Kullback-Leibler divergence [4, 5] and the Euclidean distance [7], which are discussed in the following two sections.

3.1. Symmetric Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is an information theoretically motivated measure between two probability distributions. In maximum likelihood classification where the task is to classify a signal into a class, selecting the class which minimizes the KL divergence between the observed feature distribution and the class feature distribution is the maximum likelihood class estimate [9].

The divergence between distributions $p(\mathbf{x}|\lambda_a)$ and $p(\mathbf{x}|\lambda_b)$ is defined as

$$D(p(\mathbf{x}|\lambda_a)||p(\mathbf{x}|\lambda_b)) = \int p(\mathbf{x}|\lambda_a) \log \frac{p(\mathbf{x}|\lambda_a)}{p(\mathbf{x}|\lambda_b)} d\mathbf{x}. \tag{2}$$

The divergence is not symmetric and it can therefore be symmetrized by adding the term $D(p(\mathbf{x}|\lambda_b)||p(\mathbf{x}|\lambda_a))$. When $p(\mathbf{x}|\lambda_a)$ and $p(\mathbf{x}|\lambda_b)$ are modeled using a single Gaussian distribution, the symmetric divergence can be solved in a closed form. When multiple Gaussians are used, there are several approximations for the divergence [10].

Monte-Carlo approximation calculates (2) by

$$D(p(\mathbf{x}|\lambda_a)||p(\mathbf{x}|\lambda_b)) \approx \frac{1}{R}\sum_{r=1}^{R} \log \frac{p(\mathbf{x}_r|\lambda_a)}{p(\mathbf{x}_r|\lambda_b)}, \tag{3}$$

where the samples $\mathbf{x}_r$ are drawn from distribution $p(\mathbf{x}|\lambda_a)$. An accurate approximation requires a large number of samples and is therefore computationally inefficient. Here we propose to use the samples of the observation sequence $\mathbf{A}$ that were used to train the distribution $p(\mathbf{x}|\lambda_a)$. We observe that the resulting empirical Kullback-Leibler divergence D_{emp}can be written as

$$D_{\text{emp}}(p(\mathbf{x}|\lambda_a)||p(\mathbf{x}|\lambda_b)) = \frac{1}{R} \log \frac{p(\mathbf{A}|\lambda_a)}{p(\mathbf{A}|\lambda_b)}. \tag{4}$$

Symmetric version of the above is obtained by adding the divergence $D_{\text{emp}}(p(\mathbf{x}|\lambda_b)||p(\mathbf{x}|\lambda_a))$. We assume that the lengths of the signals are equal, so that we obtain a distance measure

$$E(\mathbf{A},\mathbf{B}) = \frac{1}{R} \log \frac{p(\mathbf{A}|\lambda_a)}{p(\mathbf{A}|\lambda_b)} + \frac{1}{R} \log \frac{p(\mathbf{B}|\lambda_b)}{p(\mathbf{B}|\lambda_a)}. \tag{5}$$

The lower the above measure, the more similar $\mathbf{A}$ and $\mathbf{B}$ are. Reynolds et al. [2] denoted (5) as the symmetric Cross Entropy distance.

3.2. Euclidean distance

Helén and Virtanen used the Euclidean distance between two distributions parameterized by diagonal-covariance GMMs to measure the dissimilarity [7]. Unlike the Kullback-Leibler divergence, the exact Euclidean distance between GMMs can be calculated in a closed form. The Euclidean distance between full-covariance GMMs can be calculated by applying the method presented in [11] to calculate the correlation between individual Gaussians and combining them using the method presented in [7].

4. CROSS-LIKELIHOOD RATIO TEST

Especially in speech clustering and segmentation (see for example [1, 2, 6]) the likelihood ratio test has been used to measure the likelihood that two segments are spoken by the same speaker. The test statistic is given as

$$L(\mathbf{A},\mathbf{B}) = \frac{p(\mathbf{A}|\lambda_a)p(\mathbf{B}|\lambda_b)}{p(\mathbf{A}|\lambda_{ab})p(\mathbf{B}|\lambda_{ab})}, \tag{6}$$

where λ_{ab} is a model trained using both $\mathbf{A}$ and $\mathbf{B}$.

Instead of the above measure, we use here a modified likelihood ratio test given as

$$C(\mathbf{A},\mathbf{B}) = \frac{p(\mathbf{A}|\lambda_a)p(\mathbf{B}|\lambda_b)}{p(\mathbf{A}|\lambda_b)p(\mathbf{B}|\lambda_a)} \tag{7}$$

Here the divisor measures the likelihood that signal $\mathbf{A}$ is generated by model λ_b and signal $\mathbf{B}$ is generated by model λ_a, whereas the dividend acts as a normalization term which takes into account the complexity of both signals.

The measure (7) is used instead of (6) in our system since it is is computationally significantly less expensive to calculate because it does not require training a model for signal combinations, and it produced better results in the simulations. By taking the logarithm of (7) we end up with a measure which is identical, up to a scalar multiplier R, to the measure (5). Therefore, the lower the above measure, the more similar are $\mathbf{A}$ and $\mathbf{B}$. In (6) and (7) we can use either GMMs or HMMs to model the signals. Since the likelihood is derived similarly for both models, we do not differentiate between them but denote the model of signal $\mathbf{A}$ by λ_a and the model of signal $\mathbf{B}$ by λ_b. The cross-likelihood ratio test (7) has been previously used with GMMs in [1] and with HMMs in [12].

The measure (7) has a connection to maximum likelihood classification. If we consider each signal $\mathbf{B}$ as an individual class ω_b, the maximum likelihood classification principle classifies an observation $\mathbf{A}$ into the class having the highest conditional probability $p(\omega_b|\mathbf{A})$. If we assume that each class has the same prior probability, the likelihood of a class ω_b is $p(\mathbf{A}|\omega_b)$. The likelihood can be divided by a normalization term $p(\mathbf{A}|\omega_a)$ to obtain $p(\mathbf{A}|\omega_b)/p(\mathbf{A}|\omega_a)$. In similarity measurement we do "two-way" classification where the likelihood of signal $\mathbf{A}$ belonging to class ω_b and the likelihood of signal $\mathbf{B}$ belonging to class ω_a are multiplied. When each class ω_a is parameterized by a model λ_a, this results to the measure (7).

5. SIMULATIONS

We applied the proposed similarity measures in query-by-example, where the purpose is to retrieve signals from a database which are similar to a user-provided example. In the simulations two query methods were tested. The first was k-nearest neighbor (k-NN) query [13], which sorts the signals in order of similarity and retrieves a fixed number of most similar signals to the user. The other method was the ϵ-range query [13], where all signals having higher similarity than a predetermined threshold ϵ are retrieved.

Simulations were carried out using an audio database of speech, music, and environmental sounds. The total number of signals was 1332, each having the sampling rate 16 kHz. The signals were manually annotated into 17 classes. The classes and the number of signals in each class are listed in Table 1. The database consists of three environmental noise classes (inside car, in restaurant, and traffic), one class of drum sequences, three music classes (acoustic, electroacoustic, and symphony) taken from the RWC database, three voice classes (humming, singing, and whistling), and seven speech classes of individual speakers were taken from the CMU Arctic speech database. The database is explained more thoroughly in [14]. The classes were not used to train a classifier, but in the evaluation two signals belonging to the same class were considered to be similar. All the signals in the database are 10 seconds long random excerpts. Since the length of speech signals in the CMU Arctic database are 2-4 seconds, the signals of each speaker were concatenated to result in 10-second signals.

Each signal was divided into 46 ms frames with 50 % overlap and the following features were extracted within each frame: mel-frequency cepstral coefficients (three first coefficients, energy coefficient excluded), spectral spread, spectral centroid, spectral flux, harmonic ratio, maximum autocorrelation lag, crest factor, noise-likeness, total energy, and the variance of instantaneous power. Each feature was normalized to have zero mean and unity variance over the whole database. Principal component analysis was used to decorrelate the features.

Inside car (151)	Speaker1 (50)
In restaurant (42)	Speaker2 (47)
Traffic (38)	Speaker3 (44)
Acoustic music (264)	Speaker4 (40)
Drums (56)	Speaker5 (47)
Electroacoustic music (249)	Speaker6 (38)
Symphony music (51)	Speaker7 (50)
Humming (52)	
Singing (60)	
Whistling (53)	

Table 1: *Classes and the number of signals in each class.*

The methods tested in the simulations include the proposed GMM and HMM cross-likelihood ratio tests, KL divergence between Gaussian distributions [5], the Euclidean distance between GMMs [7], Mahalanobis distance [5], and the distance measure [14], where perceptual audio coding and lossless compression of the signals were used to obtain an information-theoretically motivated similarity measure. We used 8 Gaussian components in the GMMs, and 2 states and 8 Gaussians per each state in HMMs. The both methods used diagonal covariance matrices.

5.1. Evaluation procedure

One signal at the time was drawn from the database to serve as an example and the rest were considered as the database. The similarity measure between the example and each database signal was calculated. The total number of similarity estimations in test was therefore $S(S-1)$, where S was the number of signals in the database. After the similarities between the signals were measured, the most similar signals were retrieved either by the ϵ-range query or by the k-NN query. A retrieved signal was considered correct if it signal was from the same class as the example.

The results are presented here as a recall and precision rates. Recall means how large portion of similar signals was found from the database: $\mathrm{recall} = c/v$, c being the number of correctly retrieved signals from the database and v the sum of all the signals that would have been correct. Precision gives the portion of correct signals in all the retrieved signals: $\mathrm{precision} = c/r$, r being the total number of signals in the database that are retrieved from the class when the example is drawn from that particular class.

5.2. Results

Recall and precision for ϵ-range query with different values of ϵ are illustrated in Figure 1. When a small amount of signals is retrieved (low recall / high precision) the HMM cross-likelihood ratio test produces the best accuracy. On the other hand, when a large amount of signals is retrieved (recall $\approx$ precision), the GMM cross-likelihood ratio test provides the highest accuracy. The reason for the difference between HMMs and GMMs might be that on relatively similar signals HMMs are able to model temporal similarities and therefore to improve the accuracy, whereas on significantly different signals the use of the temporal structure only disturbs the estimation. At large precision values the compression based similarity measure works poorly. The results are different from those presented in [14] since here we used a larger number of classes and the results are calculated in a slightly different way. However, compression based method gives the best precision when the recall is very high.

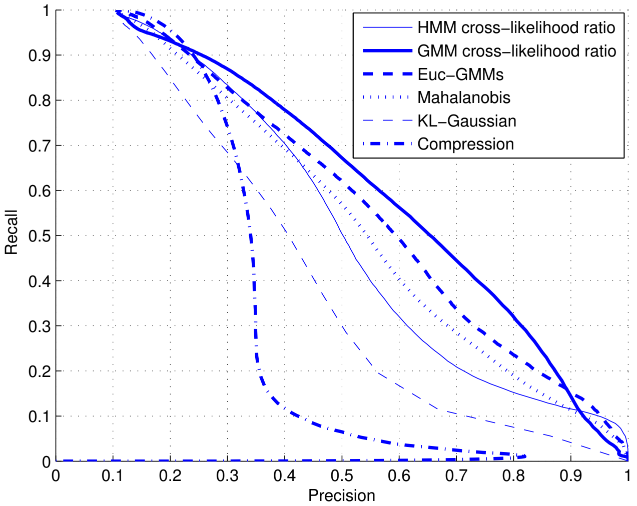

Figure 1: Results of the ϵ-range query with varying ϵ.

Table 2 presents k-NN query results when the number of retrieved signals is 10. The cross-likelihood ratio tests results in the highest precisions: 95.2% for HMMs and 94.7% for GMMs. However, almost equally well performed the Euclidean distance of GMMs with precision of 94.2%.

Similarity measure	Precision %
HMM cross-likelihood ratio	95.2
GMM cross-likelihood ratio	94.7
Euc-GMM	94.2
Mahalanobis	93.2
KL-Gaussian	91.0
Compression	77.1

Table 2: Results of the k-NN query when the number of retrieved samples was 10.

6. CONCLUSIONS

This paper proposed probabilistic model based measures for estimating the similarity of two audio signals. All the methods proposed in this paper are based on modeling the distributions of acoustic features using GMMs or HMMs. We estimate the similarity of the signals by the Kullback-Leibler divergence between GMMs, which is approximated by evaluating the distributions at points of the observed feature vectors. We also apply a cross-likelihood ratio test which uses the likelihood of the first signal being generated by the model of the second signal and vice versa. The performance of the proposed methods in query-by-example was tested using a database consisting of speech, music, and environmental sounds. None of the tested measures is superior in comparison to the others, but on average the proposed methods enable clearly better retrieval accuracy than the reference methods. The HMM-based likelihood ratio test produces good accuracy especially when a small number of most similar samples is retrieved, whereas the GMM-based likelihood ratio produces good results when a large number of samples is retrieved.

7. REFERENCES

[1] T. Stadelmann, “Fast and robust speaker clustering using the earth mover’s distance and mixmax models,” in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing*, Toulouse, France, 2006.

[2] D. A. Reynolds, E. Singer, B. A. Carlson, J. J. McLaughlin, G. C. O’Leary, and M. A. Zissman, “Blind clustering of speech utterances based on speaker and language characteristics,” in *Proc. of the Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998.

[3] C. Spevak and E. Favreau, “Soundspotter – a prototype system for content-based audio retrieval,” in *Proc. of Int. Conf. on Digital Audio Effects*, Hamburg, Germany, 2002.

[4] P. Moreno, P. P. Ho, and N. Vasconcelos, “A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications,” in *Proc. of Neural Information Processing Systems*, Vancouver, Canada, 2004.

[5] M. I. Mandel and D. P. W. Ellis, “Song-level features and support vector machines for music classification,” in *Proc. of Int. Conf. on Music Information Retrieval*, London, UK, 2005.

[6] S. Meignier, J.-F. Bonastre, and I. Magrin-Chagnolleau, “Speaker utterances tying among speaker segmented audio documents using hierarchical classification: Towards speaker indexing of audio databases,” in *Proc. of the Int. Conf. on Spoken Language Processing*, Denver, USA, 2000.

[7] M. Helén and T. Virtanen, “Query by example of audio signals using Euclidean distance between Gaussian mixture models,” in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing*, Honolulu, USA, 2007.

[8] J. Foote, “A similarity measure for automatic audio classification,” in *Proc. of AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, Palo Alto, USA, 1997.

[9] N. Vasconcelos, “On the efficient evaluation of probabilistic similarity functions for image retrieval,” *IEEE Transactions on Information Theory*, vol. 50, no. 7, 2004.

[10] J. R. Hershey and P. A. Olsen, “Approximating the Kullback Leibler divergence between Gaussian mixture models,” in *Proc. of IEEE International Conference on Audio, Speech and Signal Processing*, Honolulu, USA, 2007.

[11] P. Ahrendt, “The multivariate Gaussian probability distribution,” IMM, Technical University of Denmark, Tech. Rep., 2005.

[12] J. Yin and Q. Yang, “Integrating hidden Markov models and spectral analysis for sensory time series clustering,” in *Proc. of IEEE Int. Conf. on Data Mining*, Houston, USA, 2005.

[13] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi, “Approximate nearest neighbor searching in multimedia databases,” in *Proc. 17th IEEE International Conference on Data Engineering (ICDE)*, Heidelberg, Germany, 2001.

[14] M. Helén and T. Virtanen, “A similarity measure for audio query by example based on perceptual coding and compression,” in *Proc. Int. Conf. on Digital Audio Effects*, Bordeaux, France, Sept. 2007.