

FAST QUERY BY EXAMPLE OF ENVIRONMENTAL SOUNDS VIA ROBUST AND EFFICIENT CLUSTER-BASED INDEXING

Jiachen Xue, Gordon Wichern, Harvey Thornburg, and Andreas Spanias

Arts, Media, and Engineering
Arizona State University
Tempe, AZ 85281 USA

{jcxue, Gordon.Wichern, Harvey.Thornburg, spanias}@asu.edu

ABSTRACT

There has been much recent progress in the technical infrastructure necessary to continuously characterize and archive all sounds, or more precisely auditory streams, that occur within a given space or human life. Efficient and intuitive access, however, remains a considerable challenge. In specifically musical domains, i.e., melody retrieval, query-by-example (QBE) has found considerable success in accessing music that matches a specific query. We propose an extension of the QBE paradigm to the broad class of natural and environmental sounds, which occur frequently in continuous recordings. We explore several cluster-based indexing approaches, namely non-negative matrix factorization (NMF) and spectral clustering to efficiently organize and quickly retrieve archived audio using the QBE paradigm. Experiments on a test database compare the performance of the different clustering algorithms in terms of recall, precision, and computational complexity. Initial results indicate significant improvements over both exhaustive search schemes and traditional K-means clustering, and excellent overall performance in the example-based retrieval of environmental sounds.

Index Terms— Acoustic signal analysis, Database query processing, Clustering methods, Hidden Markov models

1. INTRODUCTION

Consistent improvements in audio processing, storage, and recording technology along with decreases in cost for the necessary hardware, currently allow for continuous, high-fidelity archival of all sounds, or more precisely *auditory streams* [1], occurring within a given space or human life [2, 3, 4]. These comprehensive audio histories allow the format of the archived recordings to dictate what events are important, as opposed to preconceived decisions controlling what is recorded. Thus, *continuous, comprehensive archival* (CCA) often provides a more holistic picture of the auditory scene by allowing the context of an audio event to be retrieved along with the event of interest. Although there are many advantages to CCA, its usefulness is often limited in practice due to the prohibitive size of the continuously growing archive. For this reason, robust, intuitive, and efficient audio information retrieval strategies are needed.

In order to provide intuitive retrieval, the query by example (QBE) paradigm is often with much success in navigating large music databases using melody content in query by humming (QBH) systems. Unfortunately, QBH methods are rather specialized to *musical* sounds where higher-level features such as melody and rhythm are salient. Audio fingerprinting techniques have achieved much success in seeking “exact matches” for databases of both music [5]

and more general audio [6], with relatively low search complexity and without utilizing high-level melodic information. However, these techniques require the user to have a specific, well-formed example of what they seek, which is not always an accurate assumption. An alternative approach consists of construction of dynamic probabilistic models for all sounds in a database, whose models are evaluated and ranked in terms of likelihood for a specific query. This allows for explicit compensation for imperfections and uncertainties in user queries [7, 8], while also providing accurate retrieval of sounds that are perceptually similar (inexact matches) to the query. For all audio information retrieval algorithms discussed up to this point, search complexity increases dramatically as the size of the database grows. Assuming that a given query is relevant to only a small percentage of sounds in the database, then clearly search time can be greatly reduced by recursively dividing the audio archive into clusters of perceptually similar sounds, and only searching those clusters that are perceptually related to the given query.

This paper provides an extension of our previous work in developing a flexible, distortion-aware QBE system for natural and environmental sounds [8]. Specifically, we present a novel semi-metric for calculating a distance matrix from the probabilistic models that index all archived sound files in a database. Two clustering methods, namely, Non-negative matrix factorization and spectral clustering, are applied to this distance matrix to partition the audio database for efficient search. Non-negative matrix factorization (NMF), decomposes the distance matrix into both bases and encoding coefficient matrices, where all elements of both factored matrices are constrained to be non-negative. Spectral clustering, transforms the distance matrix into a scaled affinity matrix and uses the dominant eigenvectors of this affinity matrix to cluster the data.

In describing our approach we begin with a brief review of the extraction of low-level feature trajectories, which are specifically tailored to natural sound environments in Section 2. As discussed in Section 3, each sound in the archive is indexed with a hidden Markov model (HMM) that incorporates *general trends* in feature trajectories (constant(low/high), up, down or fluctuating) and is robust to time-warping distortion. Once presented with a query, our system extracts feature trajectories from that query, and retrieves sounds in ranked order in terms of likelihood as their corresponding HMM is evaluated using the query features as observations. Algorithms that improve retrieval speed by clustering all sounds in the database into perceptually similar groups are discussed in Section 4. The results of Section 5 show that clustering schemes based on matrix factorization compare favorably with brute force retrieval and traditional K-means clustering in terms of recall, precision, and computational complexity. Additionally, our preliminary results demonstrate that our query

by example architecture is successful for the difficult classes of natural and environmental sounds.

2. FEATURE EXTRACTION

The audio features used to characterize the dynamic trajectories of both the queried and archived sound files were shown in [2] to represent a large variety of sounds without specifically assuming particular categories, (e.g., speech, music). Due to the diversity of sounds under consideration, we have found it necessary to calculate features at time scales from 40ms (short-term) to one second (long-term). In our study we use five short-term features: *RMS level*, Bark-weighted *spectral centroid*, *spectral sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over the short-time Fourier Transform (STFT) magnitude spectrum), *transient index* (the ℓ^2 norm of the difference of Mel frequency cepstral coefficients (MFCC's) between consecutive frames), and *harmonicity* (a probabilistic measure of whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure). The long-term feature *temporal sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one second interval), rounds out our feature set. A detailed description of how all features are computed is given in [2].

Short-term features are computed either directly from the windowed time series data or via STFT using overlapping 40ms Hamming windows hopped every 20ms. Long-term features are computed using a sliding window to combine the data from 49 of the 40ms windows. Using 98% overlap for the sliding window, (i.e., slide in 20ms steps), both long and short-term features remain synchronous. Once the features for a given sound file are computed, each trajectory is then pre-filtered using a fourth order Savitsky-Golay smoother [9], which returns not only the filtered trajectory, but also an estimate of its derivative.

3. LIKELIHOOD-BASED RETRIEVAL

Letting $t \in 1:T$ be the time index of the audio frame, for a query of length T , and $i \in 1:P$ be the feature index ($P = 6$ in our experiments), we define $Y_t^{(i)} = [x_t^{(i)}, \dot{x}_t^{(i)}]'$, as the observed feature vector at time t . Here, $x_t^{(i)}$ denotes the inherent feature value and $\dot{x}_t^{(i)}$ its derivative. We assume that all features are statistically independent, i.e.,

$$P(Y_{1:T}^{(1:P)} | \lambda^{(1:P)}(n)) = \prod_{i=1}^P P(Y_{1:T}^{(i)} | \lambda^{(i)}(n)) \quad (1)$$

where $Y_{1:T}^{(1:P)}$ are the observed features from the sound query, n is the index of the archived sound $n \in 1:N$ in a database of N sounds, and $\lambda^{(i)}(n)$ is a HMM estimated from the i th feature trajectory of archived sound n . Details on the estimation of $\lambda^{(i)}(n)$ and computation of (1) using a HMM is described in detail in [8] and summarized below.

3.1. Probabilistic Model Construction

A HMM is created by fitting constant, linear, and parabolic least squares (LS) polynomials to each smoothed feature trajectory, $Y_{1:T}^{(i)}$. These polynomial fits entale the general trends in the feature trajectories (constant(low/high), up, down or fluctuating) that the user is likely to remember. The trend type is obtained through the order of the polynomial fit, which is determined via the Akaike information criterion [10]. Next, we define the discrete hidden state

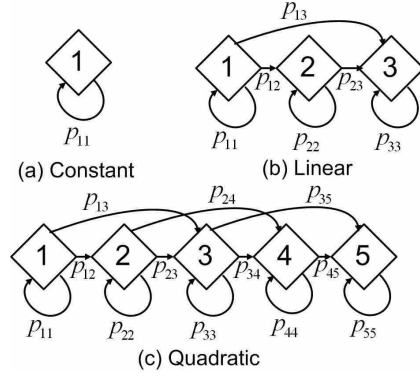


Fig. 1. Markov transition diagrams for $S_t^{(i)}$ under the three possible polynomial fits of the feature trajectories.

(mode) $S_t^{(i)}$ whose Markov transition diagrams for constant, linear, and quadratic fits, are shown in Figures 1(a), (b), and (c), respectively. The states of the HMM represent equally spaced sample points from the polynomial fit of the observation trajectory that accurately capture the overall trend of each feature. The state transition probabilities, $P(S_{t+1}^{(i)} = a | S_t^{(i)} = b) = p_{ba}$, $a, b \in 1:5$ (Figure 1) are assumed to originate from a Poisson process, where the expected time a sound remains at any one value of $S_t^{(i)}$ is dependent on the length of the sound and the shape of the polynomial fit. In order to use a HMM framework to compute (1), we let the emission probability distribution be $P(Y_t^{(i)} | S_t^{(i)}) = \mathcal{N}(\mu(S_t^{(i)}), \Sigma^{(i)})$ with $\mu(S_t^{(i)} = a)$ being the value of the fitted curve when $S_t^{(i)} = a$, while covariance matrix $\Sigma^{(i)}$ is estimated from the residuals of the observation trajectory and the optimal polynomial fit. The prior $P(S_1^{(i)} = 1) = 1$ is chosen so that likelihoods are always computed assuming the observation starts at the beginning of its trajectory.

4. LIKELIHOOD-BASED CLUSTERING

The observed feature values $Y_t^{(i)}$ are not used to cluster perceptually related sounds. Instead we construct a new discriminative space using the log-likelihood values obtained by evaluating the HMM of each sound using the feature trajectories from all other archived sounds. The clustering procedure first obtains the log-likelihood values $L(i, j) = \log P[Y_{1:T}^{(1:P)}(i) | \lambda^{(1:P)}(j)] = \sum_{k=1}^P \log P[Y_{1:T}^{(k)}(i) | \lambda^{(k)}(j)]$ for $i, j \in 1:N$, by computing the likelihood of the i th observation trajectory using the j th HMM. In the literature [11] the distance measure $D(i, j) = [L(i, j) + L(j, i)]/2$ is used to form the symmetric $N \times N$ distance matrix \mathbf{D} , whose columns are used to cluster the data. The clustering algorithms considered in this paper are generally used in a metric space (most often the n -dimensional Euclidean space \mathbb{R}^n), but this distance measure does not satisfy the triangle inequality, can be negative, and is non-distinguishable. Thus, we attempt to improve clustering performance by introducing a semi-metric and defining the elements of \mathbf{D} as

$$D(i, j) = L(i, i) + L(j, j) - L(i, j) - L(j, i). \quad (2)$$

Although the semi-metric in (2) does not satisfy the triangle inequality, its properties are: *symmetry* $D(i, j) = D(j, i)$, *non-negativity* $D(i, j) \geq 0$, and *distinguishability* $D(i, j) = 0$ iff $i = j$. Once construction of the distance matrix \mathbf{D} is complete, clustering is per-

formed on the columns of \mathbf{D} using either of the following algorithms.

4.1. Non-negative Matrix Factorization (NMF)

Given a $N \times N$ non-negative matrix \mathbf{D} , NMF [12] finds the non-negative matrix factors $\mathbf{W} \in \mathbb{R}^{N \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$ such that $\mathbf{D} \approx \mathbf{WH}$, where $K \ll N$ is typically chosen as the number of clusters [13]. As there are fewer columns in \mathbf{W} as opposed to \mathbf{D} , the columns of \mathbf{W} can be regarded as the basis vectors for \mathbf{D} . On the other hand, each column in \mathbf{H} contains the coefficients representing the degree to which each data point vector is associated with the K clusters. In this work we decompose \mathbf{D} to obtain K basis vectors capturing the perceptual qualities of each sound cluster, and each sound is represented as an additive combination of the perceptual bases. The basic steps in our clustering application of NMF are summarized as follows:

1. Select the number of clusters K .
2. Apply NMF iterative updating scheme [12, 13], which decomposes the distance matrix as $\mathbf{D} \approx \mathbf{WH}$.
3. Assign sound n to cluster k if $H_{kn} > H_{jn}$ for $j \in 1 : K$, where H_{kn} are the elements of \mathbf{H} .

4.2. Spectral Clustering

Spectral clustering refers to a broad family of algorithms where the dominant eigenvectors of a matrix containing some measure of the distance between points are used to partition data sets into clusters. In this work we use the specific spectral clustering algorithm of [14] as follows:

1. Use local scaling [14] to form the $N \times N$ affinity matrix \mathbf{A} , whose elements are obtained from the distance matrix \mathbf{D} , i.e., $A_{ij} = \exp(-D^2(i, j)/[D(i, i_M)D(j, j_M)])$. Where i_M and j_M are the M th nearest neighbor of sound i and j , respectively ($M = 10$ in this work).
2. Define \mathbf{B} to be the diagonal matrix with $B_{ii} = \sum_{j=1}^N A_{ij}$ and construct the matrix $\mathbf{J} = \mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}$.
3. Select the number of clusters K .
4. Find the K largest eigenvectors of \mathbf{J} and stack the eigenvectors in columns to form the $N \times K$ matrix \mathbf{F} .
5. Re-normalize each row of \mathbf{F} to be of unit length to form the $N \times K$ matrix \mathbf{G} , whose elements are $G_{ij} = F_{ij}/(\sum_j F_{ij}^2)^{1/2}$.
6. Treat each row of \mathbf{G} as a data point in \mathbb{R}^K and cluster via K-means or EM algorithms.
7. Assign sound i to cluster k iff the i th row of \mathbf{G} was assigned to cluster k .

Once clustering is complete, a HMM is constructed as described in Section 3.1 for each cluster using the observation trajectories of all sounds belonging to that cluster. All sounds belonging to the cluster for which the HMM exhibits the largest likelihood for a given query are returned in ranked order according to the HMM likelihood calculated using the query features as observations. As the size of the database grows, clusters can be further divided into sub-clusters by recursively applying aforementioned clustering procedures on the original cluster.

5. PRELIMINARY RESULTS

We have applied the likelihood based retrieval algorithm discussed in the previous sections to an audio database of 102 natural and environmental sounds recorded using different microphones and in various environments. All files were captured at 16bits/44.1kHz, uncompressed. The database can be loosely partitioned into five semantic categories: speech sounds, machine sounds, water sounds, whistle/animal sounds, and rhythmic sounds. We chose five example queries and had four users (adults with no known hearing impairments) rank all database sounds as relevant/non-relevant for each example query. The five example queries were: a whistle, a human imitation of a dog howl, air conditioner buzz recorded outside on a windy day, rhythmic footsteps, and a child speaking. During the user study we instructed participants to listen to each of the 102 archived sounds, and record whether or not in their opinion, each archived sound could be considered relevant to one or more of the five example queries.

In this test, we set the number of clusters to five and eight due to prior knowledge of the types of sounds in the database. Our database can be loosely partitioned into five semantic categories, but sounds in a given semantic category can sound perceptually very different from each other, e.g., rainfall, a dripping sink, and a flushing toilet are all water sounds, which have very different auditory characteristics. It is possible, however, to organize all database sounds into eight clusters where sounds belonging to the same cluster are both perceptually and semantically conjoined. The incorporation of semantic information as a way to improve audio retrieval performance as discussed in [15], remains a topic of future work.

We then benchmarked the performance of the NMF and spectral clustering algorithms, with exhaustive search and K-means algorithm using precision and recall criteria. Figures 2(a)-(d) show the recall and precision curves averaged over all queries and user rankings when the 102 sounds in the database are divided into either five or eight clusters. For the *exhaustive* search the likelihood for a specific query is evaluated with respect to the HMM for all sounds in the database, and the sounds are returned in ranked order. For the retrieval with cluster-based indexing, the HMM for each cluster is evaluated with respect to the query observation sequence, and only the most likely cluster has the likelihood of all of its member HMMs evaluated and ranked, with the sounds not in the chosen cluster returned in random order. We first show the successful performance of our overall architecture for QBE of environmental sounds. By examining the recall curves of Figures 2(a) and (b) we see that approximately 50% of the relevant sounds were retrieved among the top 10, while from the precision curves of Figures 2(c) and (d) we see that approximately 80% of the sounds ranked in the top five were considered relevant. Considering the diversity of environmental and natural sounds, these results are quite promising.

We also notice from Figures 2(a)-(d) that both recall and precision are improved for the top 10 returned sounds when the clustering is applied. This fact indicates that the clustering procedure removes sounds that are not perceptually relevant, but might contain feature trajectories similar to a given query. When only five clusters are used as shown in Figures 2(a) and 2(c), the K-means algorithm tends to perform almost as well as the exhaustive search, because two of the five clusters were far larger than the other three, and all the example queries belong to one of the large clusters. When eight clusters are used spectral clustering performs best as the chosen clusters tended to be very perceptually accurate. From Figures 2(a)-(d) we also see that spectral clustering tends to generally outperform NMF in terms of accuracy, showing the importance of the affinity map-

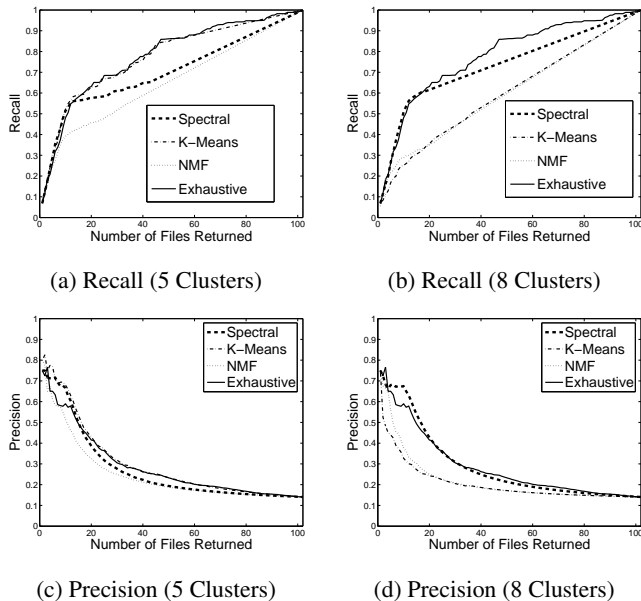


Fig. 2. Recall and precision curves averaged over five example queries and four user relevancy rankings, comparing three clustering algorithms with brute force search for five and eight clusters.

ping to clustering environmental sounds, and the difficulty of finding a low-dimensional *perceptual* basis for the broad class of natural sounds using NMF.

As a final comparison of how the clustering algorithms explored in this paper perform in a QBE system, we compare the number of HMM likelihood computations for the different algorithms and numbers of clusters in Table 1. The number of likelihood computations necessary for a given query is $\psi = |C| + K$, where $|C|$ is the number of sounds in the cluster to which the query most likely belongs. The values in this table are computed by using our 5 example queries plus all 102 sounds in the database as queries, and then computing the average number of likelihood computations for a given clustering algorithm and number of clusters. Examining Table 1 we see that all clustering algorithms far outperformed exhaustive search in terms of computational complexity and K-means tended to have the highest computational complexity for the explored clustering algorithms, because it tended to partition the space into a few very large clusters and several very small clusters. Although NMF tended to not perform as well in terms of precision and recall, it appears to find the clusters that are most computationally efficient, i.e., it partitions the database into clusters of approximately equal size.

Table 1. Average number of likelihood calculations for different indexing schemes and cluster numbers.

	Exhaustive	K-Means	Spectral Clustering	NMF
$K = 5$	102.00	37.15	32.42	28.28
$K = 8$	102.00	28.05	31.26	23.11

6. CONCLUSIONS AND FUTURE WORK

For large databases of natural and environmental sounds flexible QBE architectures that are not tailored to speech or music sounds

are necessary to provide satisfying results to human users. We show in this paper the utility and successful application of a cluster-based QBE system for improved retrieval of environmental sounds, whose complexity can remain tractable even as the number of sounds in the database grows exceedingly large. The success of our proposed scheme on environmental sounds has inspired exploration of a multi-modal (sound and gesture) QBE system as a topic of future work. We are currently developing a flexible semantic network representation which incorporates user-defined tags and relationship types. This representation is fully compatible with all clustering methods as well as the aforementioned QBE architecture.

7. REFERENCES

- [1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, 1994.
- [2] G. Wichern, H. Thornburg, B. Mechtley, A. Fink, K. Tu, and A. Spanias, "Robust multi-feature segmentation and indexing for natural and environmental sounds," in *IEEE CBMI*, Bordeaux, France, 2007.
- [3] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM CARPE*, New York, 2004.
- [4] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE ICME*, Amsterdam, The Netherlands, 2005.
- [5] Avery Wang, "The Shazam music recognition service," *Commun. ACM*, vol. 49, no. 8, pp. 44–48, 2006.
- [6] J. P. Ogle and D. P. W. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *ICASSP*, Honolulu, Hawaii, 2007.
- [7] C. Meek and W. Birmingham, "Johnny can't sing: a comprehensive error model for sung music queries," in *ISMIR*, Paris, France, 2002.
- [8] G. Wichern, J. Xue, H. Thornburg, and A. Spanias, "Distortion-aware query by example for natural sounds," in *IEEE WASPAA*, New Paltz, NY, 2007.
- [9] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, pp. 1627–1639, 1964.
- [10] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, Mar. 1974.
- [11] P. Smyth, "Clustering sequences with hidden markov models," in *Advances in Neural Information Processing Systems*, Denver, CO, 1997.
- [12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [13] W. Xu, X. Lin, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of SIGIR*, Toronto, Canada, 2003.
- [14] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, Whistler, BC, 2004.
- [15] L. Barrington, A. Chan, D. Turnbull, and G. R. G. Lanckriet, "Audio information retrieval using semantic similarity," in *ICASSP*, Honolulu, Hawaii, 2007.