

# DISTORTION-AWARE QUERY-BY-EXAMPLE FOR ENVIRONMENTAL SOUNDS

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

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

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

## ABSTRACT

There has been much recent progress in the technical infrastructure necessary to continuously characterize and archive all sounds that occur within a given space or human life. Efficient and intuitive access, however, remains a considerable challenge. In other 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. These sounds occur frequently in continuous recordings, and are often difficult for humans to imitate. We utilize a probabilistic QBE scheme that is flexible in the presence of time, level, and scale distortions along with a clustering approach to efficiently organize and retrieve the archived audio. Experiments on a test database demonstrate accurate retrieval of archived sounds, whose relevance to example queries is determined by human users.

## 1. INTRODUCTION

Recent improvements in high-capacity, high-bandwidth storage and computational auditory scene analysis are making it possible to archive all sounds which occur within a given space or human life [1, 2, 3]. Domains such as surveillance, architectural design and acoustic ecology have long been concerned with characterizing sound activity in fixed spaces [3, 4, 5], and now the emerging field of CARPE (Continuous Archival and Retrieval of Personal Experience), rooted in the ideas of Vannevar Bush [6], is finding application to domains ranging from healthcare to personal communication and even education [2, 7, 8]. All such applications can be augmented considerably via comprehensive characterization and archival of the auditory scene. Such archival is preferable to the selective archival traditionally found in these domains for essentially two reasons. First, it minimizes the effect of human preconceptions on what is archived; i.e., the structure of the space or activity itself determines what events are important. Second, continuous/comprehensive archival allows the entire *context* in which an audio event occurs to be retrieved along with that event. The user can thereby gain an awareness of how related events and their surrounding contexts become "linked" through a particular query.

Though the advantages of continuous/comprehensive archival are multifold, significant challenges remain in terms of access, which must be fast, intuitive, and supportive of contextual retrieval. To this end, query by example (QBE) is widely applied to the navigation of large music databases with melody content in the form

of *query by humming* (QBH) systems. Melody forms a convenient, perceptual *common ground* between the archived information and the information generated by the user to form queries. For more general audio archives, this common ground is much less clear. There is the trivial case of seeking "exact matches" from an archive; recent *audio fingerprinting* techniques have achieved much success [9]. However, most users may have only a vague idea of what they seek, especially at the outset. We must aim for flexible, distortion-aware QBE in the broader context of *action-based retrieval*, where users can upload "typical" examples, or choose to mimic sounds orally or by manipulating nearby objects (striking them, scratching them, and so forth).

To this end, we have developed a flexible, distortion-aware mechanism for the action-based retrieval of environmental and natural sounds. Queries are currently unimodal (sound-based); however, our framework is extensible to multi-modal cues including gesture. Our system is inspired by the extensive work in QBH (cf. [10]); however, the high-level melodic representation is replaced by a low-level feature trajectory representation consisting of features which are especially well adapted to distinguishing environmental and natural sounds. These features have been applied to the problem of environmental sound segmentation by some of the present authors [1]; we summarize them briefly in Section 2.

Before fully detailing our approach, we describe a "core method" as follows. Each sound in the archive is indexed with two dynamic Bayesian networks (DBN) built from its associated feature trajectory representation. The first DBN is a hidden Markov model (HMM) which encodes the joint distribution over all sounds the user would expect to recognize as "perceptually similar" to the given sound. This HMM models approximate feature trends (does each stay constant, go up, down, or vary in more complex ways?) and allows arbitrary distortions of the time axis, similarly to dynamic time warping models [11]. The second DBN augments the HMM in the form of a switching state-space model (SSM) which models additional uncertainties in query production mechanisms, in the form of affine distortions of feature trajectories. Both DBN's are discussed in Section 3. Once presented with a query, our system extracts feature trajectories from that query, and retrieves the  $M$  sounds with the highest likelihoods as their corresponding network is evaluated using the query features as observations.<sup>1</sup> An unfortunate aspect is the  $\mathcal{O}(N)$  complexity of retrieval where  $N$  is the number of database sounds. Instead, we have developed

<sup>1</sup>In our actual database, sounds are indexed with start/stop time and other information such as GPS location and various levels of user annotation. This information enables us to retrieve not only the sounds, but as much of the surrounding context as desired, for instance, minutes-long sound clips from the same environment which contains the given sound.

\*This material is based upon work supported by the National Science Foundation under Grant No. 0504647.

an efficient cluster-based indexing (Section 3.4) which reduces the average complexity to  $\mathcal{O}(\log N)$  in practice. Despite the approximation inherent in this scheme, results (Section 4) show that our overall method works quite well.

Because retrieval goals are dependent on human action we must take into account the perception of the end-user when evaluating our system. The results of Section 4 are based upon a user study where 102 archived sounds are labeled as relevant/non-relevant with respect to example queries exhibiting realistic distortions. The outcomes of these user tests are then averaged and used to quantitatively evaluate our QBE system in terms of the well-known precision and recall metrics.

## 2. FEATURE EXTRACTION

The audio features used to characterize the dynamic trajectories of both the queried and archived sound files were chosen to represent a large variety of sounds without specifically assuming particular categories, e.g., speech or 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  $L^\infty$  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  $L^\infty$  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 all features can be found in [1].

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, 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 recording of length  $T$ , and  $i \in 1:P$  be the feature index ( $P = 6$  in our experiments), we define  $Y_t^{(i)} = [x_t^{(i)}, \dot{x}_t^{(i)}]^T$ , 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.,

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

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  $\lambda_n^{(i)}$  is a DBN estimated from the  $i$ th feature trajectory

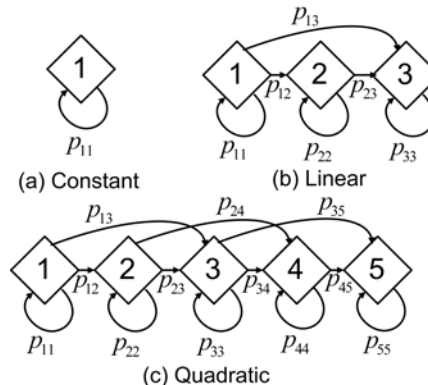


Figure 1: Markov transition diagrams for  $M_t^{(i)}$  under the three possible polynomial fits of the feature trajectories.

of archived sound  $n$ . Details on the estimation of  $\lambda_n^{(i)}$  and computation of (1) using either a HMM or the novel distortion-aware DBN will be described next.

### 3.1. Probabilistic Model Construction

The smoothed feature trajectories are used to automatically create a DBN for every archived sound file by fitting constant, linear, and parabolic least squares (LS) polynomials to each observation trajectory,  $Y_{1:T}^{(i)}$ . The Akaike information criterion is used to determine the optimal polynomial order for a given observation trajectory. The optimal fit is denoted by  $S_t^{(i)} = [z_t^{(i)}, \dot{z}_t^{(i)}]^T$ , where  $z_t^{(i)}$  is the value of the fit at time  $t$ , and  $\dot{z}_t^{(i)}$  is the derivative of the fit at time  $t$ . Next, we define the discrete hidden mode  $M_t^{(i)}$  whose Markov transition diagrams for constant, linear, and quadratic fits, are shown in Figures 1(a), (b), and (c), respectively. The possible values of  $M_t^{(i)}$  are equally-spaced sample points from the polynomial fit of the observation trajectory, e.g., Figure 2, for a quadratic fit of a harmonicity trajectory, with the dots representing the values of  $M_t^{(i)}$ . The mode transition probabilities,  $P(M_{t+1}^{(i)} = a | M_t^{(i)} = b) = p_{ba}$ ,  $a, b \in 1:5$  (Figure 1) are assumed to originate from a Poisson process, where the expected time a sound remains at any one value of  $M_t^{(i)}$  is  $1/d_a^{(i)}$ , and  $d_a^{(i)}$  is the frame difference between two consecutive values of  $M_t^{(i)}$  (e.g., Figure 2). Clearly, since  $M_t^{(i)}$  is discrete we can use a HMM framework to compute (1), where the emission probability distribution is  $P(Y_t^{(i)} | M_t^{(i)}) = \mathcal{N}(\mu(M_t^{(i)}), \Sigma^{(i)})$  with  $\mu(M_t^{(i)})$  being one of the possible values for  $M_t^{(i)}$ , while covariance matrix  $\Sigma^{(i)}$  is estimated from the residuals of the observation trajectory,  $Y_{1:T}^{(i)}$ , and the optimal LS polynomial fit  $S_{1:T}^{(i)}$ . The prior  $P(M_1^{(i)} = 1) = 1$  is chosen so that likelihoods are always computed assuming the observation starts at the beginning of its trajectory.

### 3.2. Distortion-Aware Extension

Human imperfections when trying to mimic natural sounds, as well as differences in recording conditions can cause distortions in feature trajectories. To model affine distortions in the query feature trajectories, we define the continuous-valued state vector  $V_t^{(i)} = [\alpha_t^{(i)}, \beta_t^{(i)}]^T$ , where  $\alpha_t^{(i)}$  models level distortion in the

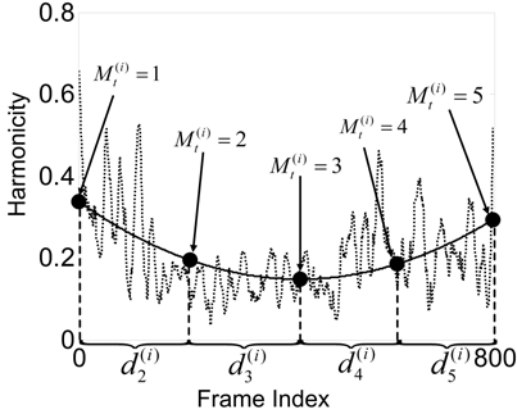


Figure 2: Example of LS quadratic fit (solid line) and corresponding discrete mode values to harmonicity trajectory (dotted line).

query, and  $\beta_t^{(i)}$  models scale distortion. The HMM can be made distortion-aware by modifying the distribution of the observation, i.e.

$$P(Y_t^{(i)} | V_t^{(i)}, M_t^{(i)}) \quad (2)$$

$$= \mathcal{N} \left( \beta_t^{(i)} \mu(M_t^{(i)}) + \begin{bmatrix} \alpha_t^{(i)} - \beta_t^{(i)} z_0^{(i)} \\ 0 \end{bmatrix}, [\beta_t^{(i)}]^2 \Sigma^{(i)} \right)$$

where  $z_0^{(i)}$  is the mean calculated over  $t$  of the polynomial fit  $z_{1:T}^{(i)}$ . In order to compute the likelihood in the distortion aware model, we no longer have a HMM, but rather a switching state space model (SSM), and must use a multiple model approach for likelihood estimation [12]. In this approach a bank of dynamic models, each matched to a specific value of discrete mode  $M_t^{(i)}$  are combined in order to infer the values of the hidden variable sequences  $V_{1:T}^{(i)}$  and  $M_{1:T}^{(i)}$ . For a given value of  $M_t^{(i)}$  the dynamic system governing the distortion-aware model is described by

$$V_t^{(i)} = V_{t-1}^{(i)} \quad (3)$$

$$Y_t^{(i)} = \beta_t^{(i)} \mu(M_t^{(i)}) + \begin{bmatrix} \alpha_t^{(i)} - \beta_t^{(i)} z_0^{(i)} \\ 0 \end{bmatrix} + \beta_t^{(i)} n_t^{(i)} \quad (4)$$

where  $n_t^{(i)} \sim \mathcal{N}(0, \Sigma^{(i)})$  is a Gaussian noise source.

### 3.3. Likelihood Computation

A fundamental problem in SSM's is that if  $M_t^{(i)}$  can take  $R$  possible values, then  $P(Y_t^{(i)} | Y_{1:t}^{(i)})$  is a mixture of  $R^t$  Gaussians, one for each possible sequence of  $M_{1:t}^{(i)}$ . In order to overcome this computational intractability we employ the Generalized Pseudo Bayesian approximation of order 2 (GPB2) [12] where the Kalman time and measurement updates have been replaced by the appropriate unscented modifications [13], due to the inherent nonlinearity of the continuous-state dependence in (4).

Given a DBN model for feature  $i$ , each term in the product of likelihoods in (1) can be factored as

$$P(Y_{1:T}^{(i)}) = \prod_{t=1}^T P(Y_{t+1}^{(i)} | Y_{1:t}^{(i)}) \quad (5)$$

where

$$P(Y_{t+1}^{(i)} | Y_{1:t}^{(i)}) = \sum_{M_t^{(i)}} \sum_{M_{t-1}^{(i)}} P(Y_t^{(i)} | M_{t-1}^{(i)}, M_t^{(i)}, Y_{1:t}^{(i)}) \quad (6)$$

$$\times P(M_t^{(i)} | M_{t-1}^{(i)}, Y_{1:t}^{(i)}) P(M_{t-1}^{(i)} | Y_{1:t}^{(i)})$$

and details on the computation of (6) are provided in [12].

### 3.4. Clustering for Efficient Retrieval

The observed feature values  $Y_t^{(i)}$  are not used to cluster perceptually related sounds, instead we attempt to find 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 following [14]. We use the HMM instead of the distortion-aware DBN for clustering, because we are interested in computing perceptual similarities between sounds, without taking into account specific distortions due to the human reproduction of those sounds. The clustering procedure first obtains the log-likelihood values  $L_{n\ell} = \log P(Y_{1:T}^{(1:P)}(n) | \lambda_\ell^{(1:P)})$  by running observation trajectory  $Y_{1:T}^{(1:P)}(n)$ , for sound  $n$  through all HMMs  $\lambda_\ell^{(1:P)}$ ,  $\ell \in 1 : N$ . We then form the  $N \times N$  similarity matrix  $\Psi$  whose entries are  $\psi_{n\ell} = (L_{n\ell} + L_{\ell n})/2$ . Clustering is then performed on the rows of  $\Psi$ .

We adopt a recursive K-means algorithm to automatically form hierarchical clusters that facilitate efficient search of large databases. The recursive K-means algorithm initially divides the entire data set into  $k$  clusters using the standard K-means algorithm, and then further sub-divides these clusters until the number of elements in any given cluster is less than a pre-defined threshold. Once recursive K-means 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. Prior to HMM estimation all archived sounds are resampled, so every sound in a given cluster contains an equal number of observations. The sub-cluster whose HMM exhibits the largest likelihood for a given query has the likelihood of all of its sounds calculated using the distortion-aware procedure described in Section 3.2, and ranks those sounds in order of greatest likelihood. Due to the clustering the overall search complexity becomes  $\mathcal{O}(\log N)$ .

## 4. 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 102 archived sounds were organized into three clusters containing 17, 21, and 64 sound files, using the algorithm described in Section 3.4.

We chose five example queries and had four users rank all 102 sounds in the database 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.

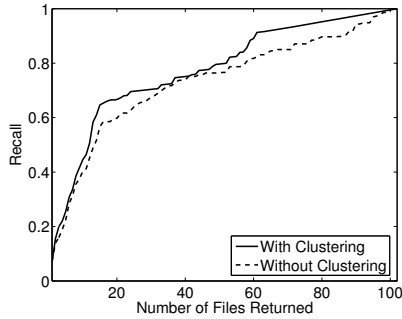


Figure 3: Recall curve averaged over five example queries and four user relevancy rankings, both with and without clustering.

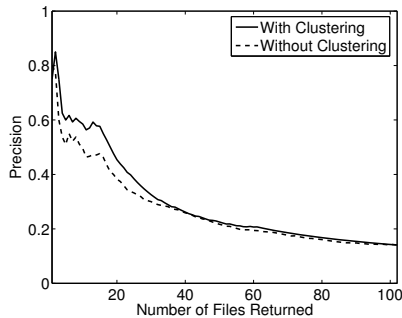


Figure 4: Precision curve averaged over five example queries and four user relevancy rankings, both with and without clustering.

We then evaluated the performance of our system using precision and recall criteria.<sup>2</sup> Figure 3 displays the average recall of our QBE system, as a function of the number of retrieved sounds, both with clustering (only the  $M$  sounds belonging to the chosen cluster are ranked, with the remaining sounds returned in random order) and without clustering (all archived sounds are returned and ranked). By examining specific points on the recall curve of Figure 3 we see that approximately 50% of the relevant sounds were retrieved among the top 10, and 70% of the relevant sounds were retrieved in the top 20. Similar to Figure 3, Figure 4 displays the average precision as a function of the number of retrieved sounds. Examining points on the curve of Figure 4 we see that approximately 80% of the sounds ranked in the top five and 60% of those ranked in the top 20 were considered relevant.

As a final comment on Figures 3 and 4 we see that both recall and precision were improved by use of the clustering algorithm as candidate sounds that were not perceptually relevant, but might contain feature trajectories similar to the query, are removed from consideration by the clustering procedure.

## 5. 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 of a distortion-aware QBE system

<sup>2</sup>Recall is defined as the number of relevant sounds returned divided by the total number of relevant sounds in the database. Precision is the number of relevant sounds returned divided by the number of returned sounds. See [15] for further discussions.

that accounts for the imperfections of queries composed of human actions, and provides a framework for efficient, low-latency retrieval even as the number of sounds in the database grows exceedingly large. The extension of our proposed scheme to multimodal (sound and gesture) queries, exploration of different indexing schemes, incorporation of temporal expectancy models [16], and the investigation of feature mapping between the human voice and sounds that it cannot accurately imitate [17] are all possibilities to improve the experience of the end user for action-based retrieval in the future.

## 6. REFERENCES

- [1] 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.
- [2] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM CARPE*, New York, 2004.
- [3] 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.
- [4] R. M. Schafer, *The Soundscape*. Destiny Books, 1968.
- [5] P. Hedfors and P. G. Berg, "The sounds of two landscape settings: auditory concepts for physical planning and design," *Landscape Research*, vol. 28, pp. 245–263, 2003.
- [6] V. Bush, "As we may think," *Atlantic Monthly*, July 1945.
- [7] J. Gemmell, G. Bell, and R. Lueder, "MyLifeBits: a personal database for everything," *Commun. ACM*, vol. 49, no. 1, pp. 88–95, 2006.
- [8] D. Birchfield, T. Ciuffo, and H. Thornburg, "Sound and interaction in K-12 mediated education," in *ICMC*, New Orleans, LA, USA, 2006.
- [9] A. Wang, "The Shazam music recognition service," *Commun. ACM*, vol. 49, no. 8, pp. 44–48, 2006.
- [10] R. B. Dannenberg and N. Hu, "Understanding search performance in query-by-humming systems," in *Int. Symposium on Music Information Retrieval (ISMIR)*, 2004.
- [11] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *IEEE WASPAA*, New Paltz, NY, 2003.
- [12] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*. London: Academic Press, 1988.
- [13] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, 2004.
- [14] P. Smyth, "Clustering sequences with hidden markov models," in *Advances in Neural Information Processing Systems*, Denver, CO, 1997.
- [15] C. J. V. Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.
- [16] H. Thornburg, D. Swaminathan, T. Ingalls, and R. Leistikow, "Joint segmentation and temporal structure inference for partially observed event sequences," in *IEEE MMSP*, Victoria, BC, 2006.
- [17] K. Dobson, B. Whitman, and D. P. W. Ellis, "Learning auditory models of machine voices," in *IEEE WASPAA*, New Paltz, NY, 2005.