

# Incorporating Semantics into a Query-by-Example Paradigm for Audio Information Retrieval

[Extended Abstract (Poster Submission)]

Gordon Wichern  
Arts, Media and Engineering  
Arizona State University  
Tempe, AZ  
Gordon.Wichern@asu.edu

Harvey Thornburg  
Arts, Media and Engineering  
Arizona State University  
Tempe, AZ

Andreas Spanias  
Department of Electrical  
Engineering  
Arizona State University  
Tempe, AZ  
spanias@asu.edu

## 1. INTRODUCTION

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 general audio information retrieval, including 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. Additionally, we aim for flexible, distortion-aware QBE in the broader context of *action-based retrieval*, where users can upload "typical" examples, mimic sounds orally, manipulate nearby objects (strike them, scratch them, and so forth), or use semantic queries and relationships to retrieve audio or other multimedia objects.

## 2. EXAMPLE-BASED AUDIO RETRIEVAL

For the audio information retrieval problem, the objective is that users should be able to retrieve all sound events of interest in an intuitive and efficient manner, without too many false positives. By "intuitive" we mean that the user can quickly form the query; by "efficient" we mean that the system can quickly execute the retrieval. Consider, for instance, a security application involving continuous monitoring of a warehouse. One may wish to obtain all sounds of a particular type, for instance speech, to learn what was

discussed and when. Such an application is generally *recall-driven*<sup>1</sup> as it is important not to miss any relevant sounds - the cost of a false positive is limited only by the total time it takes to listen to all of the retrieved sounds. On the other hand, another user may be interested in when all of the footsteps sounds occur, to understand when people enter or exit corridors and meeting places. This application is more *precision-driven*, as too many false positives might make it unclear when footsteps actually occur.

Due to this variety of applications, we seek a flexible strategy that allows us to navigate the recall-precision tradeoff by varying retrieval size. A convenient strategy is *query-by-example* (QBE), where users input recordings they consider similar to the desired retrieval sounds. Users can either upload the query from a file or present it orally. Oral query is quite prevalent in melody retrieval in the form of query-by-humming (QBH) ([2]; among others), where the sound objects to be retrieved concern melodies. During retrieval, the  $N$  database melodies considered "most similar" to the query are obtained.

To consider QBE in more general terms, we first need some definitions. Given  $M$  database sounds (e.g. melodic recordings), let  $X^{(i)}$  denote the  $i^{\text{th}}$  sound, and  $F^{(i)}$  denote the corresponding feature set (e.g. melody) indexed along with the  $X^{(i)}$ . Let  $Y$  denote the input sound (e.g. the recording of the user humming a melody), and  $G$  a corresponding feature set that summarizes information in  $Y$  relevant for the retrieval process. For example, the QBH approach of Durey and Clements [2] takes  $G$  to be a time-series of frame-wise pitch estimates. Assuming the intended retrieval object is in the database, the goal is to maximize a *similarity*  $J(F^{(i)}, G)$  with respect to  $i$  that increases the more similar are  $X^{(i)}$  and  $Y$ . To determine an effective similarity measure, we first revisit the original precision and recall objectives. For retrieval size = 1, and assuming the intended object exists in the database, it is easily shown that both expected precision/recall objectives reduce to the probability of retrieving the correct melody. An optimal retrieval system will maximize this probability, which is equivalent to maximizing the posterior  $P(F^i|G)$ . With the additional as-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

<sup>1</sup>*Recall* (number of desired sounds retrieved/number of desired sounds in the database) and *precision* (number of desired sounds/number of retrieved sounds) are standard measures of retrieval performance; see [4] for further information.

sumption that all recordings are equally likely (i.e.,  $P(F^{(i)})$  is uniform), the best choice of  $J(F^{(i)}, G)$  reduces to the likelihood  $P(G|F^{(i)})$ . That is, we apply the *maximum likelihood* criterion, retrieving sound  $X^{(i*)}$ , where

$$i^* = \operatorname{argmax}_{i \in 1:M} P(G|F^{(i)}) \quad (1)$$

For retrieval size  $R > 1$ , (1) can be extended to obtain the  $X^i$  with the  $R$  greatest likelihoods  $P(G|F^{(i)})$ . In this way the precision-recall tradeoff can be navigated by varying  $R$ . Not surprisingly, a vast number of QBH systems [2] are likelihood based.

While we can consider the likelihood,  $P(G|F^{(i)})$ , as a measure of similarity between query and database sounds, we can interpret it more accurately as a model of the user’s query behavior, since  $P(G|F^{(i)})$  represents the distribution of input features  $G$  that a user will produce to retrieve a sound indexed with features  $F^{(i)}$ . Query behavior is certainly *influenced* by the perceptual similarity between query and database sounds. However, the resources available to perform the query will affect how the query is performed. File-based queries are limited by the examples the user has on hand, or can easily obtain, while oral queries are limited by the range of the user’s vocal apparatus. Especially when retrieving “non-vocal” sounds such as door slams, gunshots, explosions, and so forth, the query sound may be perceptually far more dissimilar to the target in the case of oral query, as compared with file-based query - the presence of onomatopoeic words in almost every human language [1] attests to the diversity of ways humans have learned to convey to one another the occurrence of a given sound.

As such, the development of robust query likelihood models for natural sounds must account not only for the perceptual similarity between query and target sounds; it must accurately model the entire range of query behavior given available resources and the situated context of the query action. To this end, we have developed a flexible likelihood model for file-based query of environmental sounds, which models query behavior and can be readily adapted for oral query. This model explicitly addresses the types of distortions users are likely to produce in the query process. Similar to likelihood-based QBH, our query feature set is a *time series*;  $G = G_{1:T}$ , where each  $G_t$  is a vector consisting of the six features described in [5], and  $t$  is the frame index. The database feature set  $F^{(i)}$  also represents a time series in the same feature vectors; however each individual feature trajectory is modeled as a *general trend*, consisting of a zeroth, first, or second-order polynomial fit. These fits suffice to encode whether the feature is constant (high or low), increasing/decreasing, or exhibiting more complex (up  $\rightarrow$  down; down  $\rightarrow$  up) behavior. Our model,  $P(G|F^{(i)})$ , is a DBN that addresses the following types of query distortion: a) time-warping distortion - queries can be long or short, and the user may be relatively insensitive to the time scale over which features evolve; b) additional tolerance due to the limited number of query examples.

### 3. INCORPORATION OF SEMANTIC INFORMATION

Although, the query-by-example framework has several advantages, like any multimedia information retrieval strategy it also has its shortcomings. Up to this point we have

only considered content-based retrieval, which tries to retrieve sounds based on their perceptual similarity to the query. Many multimedia objects also contain explicit semantic information or connotations that connect them, despite the differences of their sonic qualities. A sound of a ship, for example, may not contain qualities that allow it to be retrieved when the query is an ocean wave, despite the intuitive closeness of these two sounds.

Semantic information can be represented through an ontology where users can manually provide information linking sounds to each other or to other multimedia and concepts. Computationally, a metric such as shortest path distance can be used to measure distances between concepts. These distances can then be combined, using tuned weights, with the feature-space affinities. Additionally, external ontologies such as ConceptNet [3] can be used to situate sparse user-provided information into a larger semantic structure, while also tailoring performance for specific user communities. One strength of including semantic information in this fashion, as opposed to only relying on tags, is that edges representing different types of conceptual linkages can be weighted differently. For example, if an external ontology is used in indexing distances between sounds, the bias of this community knowledge can be adjusted. Semantic information can also assist in generalization of the audio information retrieval process. In cases where the retrieval process is not only oral, for example, semantic relations between sounds and segments of other activity, such as physical gesture or text-based query, can be used to create a more robust action-based retrieval system.

The key to extending our multimedia information retrieval algorithm to additional modalities such as gesture, text, images, etc. is the concept of a probabilistic template. A dynamic example for sounds was described above, but the probabilistic template concept extends to any method that compares two multimedia objects in order to compute the likelihood  $P(G|F^{(i)})$  where  $G$  contains features summarizing the content of the query, and  $F^{(i)}$  is the template summarizing the  $i$ th database object (or group of objects). These templates can then be linked to semantic concepts, which can in turn be linked to additional multimedia databases. An example of linking multiple multimedia databases via semantic concepts is shown in Figure 1. The directed edges connecting the query, templates, and semantic concepts have associated weights, where the smaller the weight the more related are the nodes it connects. Thus, we can use the sum of the weights on the path between any two nodes in the network to be a measure of their similarity.

In the example of Figure 1 human movement is used as the query, and is matched to a series of probabilistic gesture templates, where the weight of the edge connecting the query to template  $i$  is  $-\log P(G|F^{(i)})$ , where  $G$  and  $F^{(i)}$  summarize the content of the query and  $i$ th database template, respectively. The weights of the edges connecting the templates and the semantic concepts can be set by the interaction designer, or learned over time from user data. In Figure 1 a database of sounds are connected to the right of the semantic concepts, and the output of this system would be a distribution over all  $N$  sounds in the database. This distribution can be determined by computing the shortest path from the gesture query to each sound, and then normalizing so the probability of sound  $i$  is  $p(\text{sound}_i) = \exp(-\text{path}_i) / \sum_j^N \exp(-\text{path}_j)$ , where  $\text{path}_i$  is

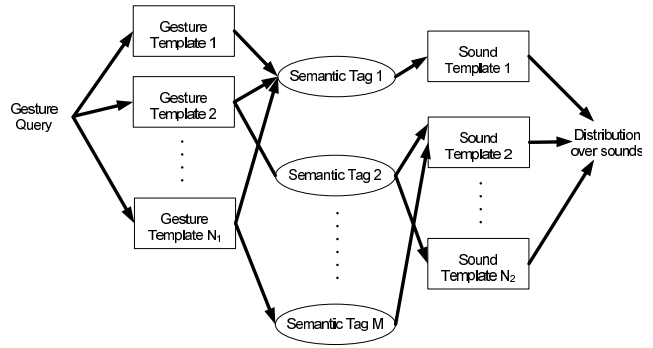
the shortest path between the query and the node representing sound  $i$ . Here a path is defined as a sequence of nodes connected by directed edges, where each node is visited no more than one time. Additionally, if there are multiple paths between the query and the node corresponding to sound  $i$ , then  $path_i$  can represent the geometric mean of all paths, or if that is computationally infeasible, the geometric mean of the  $K$  shortest paths.

It is also possible to use a semantic network similar to the one shown in Figure 1 for annotation and text query. In an annotation task for this example, the sound nodes and the edges connecting them to the semantic concept nodes would be removed from the network. Thus, a gesture query would provide a distribution over semantic concepts, which could automatically describe or annotate the gesture. For this example, a system using semantic (text) query could be obtained by removing the gesture template nodes and connecting the query directly to the semantic nodes. A text query could then be used to compute the distribution over sounds. Additionally, a traditional content-based query by example system also exists in the semantic network of Figure 1 if the gesture query is used to output a distribution over the gesture templates.

In traditional multimedia retrieval applications the distribution over multimedia objects output by the system would be a ranked list in order of decreasing probability. In this work we propose treating multimedia information retrieval as the basis for a media feedback engine. For example, a user moving in a space can have their movements serve as the gesture input to the semantic network of Figure 1. The feedback the user receives could then be a soundscape composed of combining the  $N$  sounds in the database in a manner consistent with their distribution for the given gesture query. This method of feedback allows for *probabilistic interaction*, and can help make multimedia environments more engaging for the user. Traditional multimedia interactions tend to be rather causal, e.g., “when I move my arm up, the sound gets louder and the red light turns on.” Probabilistic interaction should allow the user to comprehend general feedback concepts, while also providing an interaction that remains interesting and novel after repeated use.

We now provide two specific examples where we utilize the semantic network retrieval system just discussed. The first application allows for oral retrieval of audio files. As discussed previously, oral retrieval is very limited by the user’s vocal apparatus, therefore content-based oral retrieval of sounds not generated by the human voice is very difficult. We propose to help tackle the limitations of the human voice using semantics. We first begin by choosing a vocabulary of semantic concepts, and having users choose whether or not each word in the vocabulary is relevant to each sound file in the database. We can then use these results to learn the weights of the edges connecting the semantic nodes to the database sounds. We can then create human voice templates for each semantic concept by having users make sounds that they feel correspond to each concept in the vocabulary. Once all weights have been learned, a user can query the database with their voice and retrieve the ranked list of sounds in the database. By accounting for semantic information as well as the limitations in the human voice, we hope to provide a more intuitive retrieval experience.

A second example, takes place in a full body gaming environment where user gestures and actions correspond to



**Figure 1: Semantic network example of a gesture driven media feedback engine.**

sustainable water usage in the desert. A database of sounds corresponding to the desert environment and human water usage is used as the media feedback engine. Using a semantic network similar to that of Figure 1, a probabilistic interaction allows users to control the soundscape in the space through their gestures.

## 4. EVALUATION

In order to accurately evaluate a multimedia information retrieval algorithm like the one presented here, human users must be included. Developed algorithms will be evaluated in terms of traditional information retrieval metrics such as precision, recall and mean average precision, while also utilizing user data to learn semantic weights and evaluate the effectiveness of probabilistic interactions.

We can then use these results to learn the weights of the edges connecting the semantic nodes to the database sounds by the following process. First, one desired output distribution over all database sounds per semantic concept will be estimated from the relevance information using direct maximum-likelihood estimation if the database is small, or an appropriate nonparametric method if the database is large. Next, we will form an objective function which is the average Kullback-Leibler information between each desired output distribution and each actual output distribution given the single semantic concept as a query. Finally, we will use an appropriate nonlinear optimization technique based on gradient descent to learn the weights that minimize the objective function.

## 5. REFERENCES

- [1] N. Burton-Roberts, P. Carr, and G. J. Docherty. *Phonological Knowledge: Conceptual and Empirical Issues*. Oxford University Press, 2000.
- [2] A. S. Durey and M. A. Clements. Melody spotting using hidden markov models. In *ISMIR*, Bloomington, Indiana, 2001.
- [3] H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211–226, 2004.
- [4] C. J. V. Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
- [5] 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.