Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 192363, 11 pages
doi:10.1155/2010/192363

Research Article

An Ontological Framework for Retrieving Environmental Sounds Using Semantics and Acoustic Content

Gordon Wichern, Brandon Mechtley, Alex Fink, Harvey Thornburg, and Andreas Spanias

Arts, Media, and Engineering and Electrical Engineering Departments, Arizona State University, Tempe, AZ 85282, USA

Correspondence should be addressed to Gordon Wichern, gordon.wichern@asu.edu

Received 1 March 2010; Accepted 19 October 2010

Academic Editor: Andrea Valle

Organizing a database of user-contributed environmental sound recordings allows sound files to be linked not only by the semantic tags and labels applied to them, but also to other sounds with similar acoustic characteristics. Of paramount importance in navigating these databases are the problems of retrieving similar sounds using text- or sound-based queries, and automatically annotating unlabeled sounds. We propose an integrated system, which can be used for text-based retrieval of unlabeled audio, content-based query-by-example, and automatic annotation of unlabeled sound files. To this end, we introduce an ontological framework where sounds are connected to each other based on the similarity between acoustic features specifically adapted to environmental sounds, while semantic tags and sounds are connected through link weights that are optimized based on user-provided tags. Furthermore, tags are linked to each other through a measure of semantic similarity, which allows for efficient incorporation of out-of-vocabulary tags, that is, tags that do not yet exist in the database. Results on two freely available databases of environmental sounds contributed and labeled by nonexpert users demonstrate effective recall, precision, and average precision scores for both the text-based retrieval and annotation tasks.

1. Introduction

With the advent of mobile computing, it is currently possible to record any sound event of interest using the microphone onboard a smartphone, and immediately upload the audio clip to a central server. Once uploaded, an online community can rate, describe, and reuse the recording appending social information to the acoustic content. This kind of user-contributed audio archive presents many advantages including open access, low cost entry points for aspiring contributors, and community filtering to remove inappropriate content. The challenge in using these archives is overcoming the "data deluge" that makes retrieving specific recordings from a large database difficult.

The content-based query-by-example (QBE) technique where users query with sound recordings they consider acoustically similar to those they hope to retrieve has achieved much success for both music [1] and environmental sounds [2]. Additionally, content-based QBE is inherently unsupervised as no labels are required to rank sounds in terms of their similarity to the query (although relevancy labels are required for formal evaluation). Unfortunately, even if suitable recordings are available they might still be insufficient to retrieve certain environmental sounds. For example, suppose a user wants to retrieve all of the "water" sounds from a given database. As sounds related to water are extremely diverse in terms of acoustic content (e.g., rain drops, a flushing toilet, the call of a waterfowl, etc.), QBE is inefficient when compared to the simple text-based query "water." Moreover, it is often the case that users do not have example recordings on hand, and in these cases text-based semantic queries are often more appropriate.

Assuming the sound files in the archive do not have textual metadata, a text-based retrieval system must relate sound files to text descriptions. Techniques that connect acoustic content to semantic concepts present an additional challenge, in that learning the parameters of the retrieval system becomes a supervised learning problem as each training set sound file must have semantic labels for parameter learning. Collecting these labels has become its own research problem leading to the development of social games for collecting the metadata that describes music [3, 4].

Most previous systems for retrieving sound files using text queries, use a supervised multicategory learning approach where a classifier is trained for each semantic concept in the vocabulary. For example, in [5] semantic words are connected to audio features through hierarchical clusters. Automatic record reviews of music are obtained in [6] by using acoustic content to train a one versus all discriminative classifier for each semantic concept in the vocabulary. An alternative generative approach that was successfully applied to the annotation and retrieval of music and sound effects [7] consists of learning a Gaussian mixture model (GMM) for each concept. In [8] support vector machine (SVM) classifiers are trained for semantic and onomatopoeia labels when each sound file is represented as a mixture of hidden acoustic topics. A large-scale comparison of discriminative and generative classification approaches for text-based retrieval of general audio on the Internet was presented in [9].

One drawback of the multiclass learning approach is its inability to handle semantic concepts that are not included in the training set without an additional training phase. By not explicitly leveraging the semantic similarity between concepts, the classifiers might miss important connections. For example, if the words "purr" and "meow" are never used as labels for the same sound, the retrieval system cannot model the information that these sounds may have been emitted from the same physical source (a cat), even though they are widely separated in the acoustic feature space. Furthermore, if none of these sounds contain the tag "kitty" a user who queries with this out of vocabulary tag might not receive any results, even though several cat/kitty sounds exist in the database.

In an attempt to overcome these drawbacks we use a taxonomic approach similar to that of [10, 11] where unlabeled sounds are annotated with the semantic concepts belonging to their nearest neighbor in an acoustic feature space, and WordNet [12, 13] is used to extend the semantics. We aim to enhance this approach by introducing an ontological framework where sounds are linked to each other through a measure of acoustic content similarity, semantic concepts (tags) are linked to each other through a similarity metric based on the WordNet ontology, and sounds are linked to tags based on descriptions from a user community.

We refer to this collection of linked concepts and sounds as a *hybrid (content/semantic) network* [14, 15] that possesses the ability to handle two query modalities. When queries are sound files the system can be used for automatic *annotation* or "autotagging", which describes a sound file based on its audio content and provides suggested tags for use as traditional metadata. When queries are concepts they can be used for *text-based retrieval* where a ranked list of unlabeled sounds that are most relevant to the query concept is returned. Moreover, queries or new sounds/concepts can be efficiently connected to the network, as long as they can be linked either perceptually if sound based, or lexically if word based.

In describing our approach, we begin with a formal definition of the related problems of automatic annotation and text-based retrieval of unlabeled audio, followed by the introduction of our ontological framework solution in Section 2. The proposed hybrid network architecture outputs a distribution over sounds given a concept query (text-based retrieval) or a distribution over concepts given a sound query (annotation). The output distribution is determined from the shortest path distance between the query and all output nodes (either sounds or concepts) of interest. The main challenge of the hybrid network architecture is computing the link weights. Section 3 describes an approach to determine the link weights connecting sounds to other sounds based on a measure of acoustic content similarity, while Section 4 details how link weights between semantic concepts are calculated using a WordNet similarity metric. It is these link weights and similarity metrics that allow queries or new sounds/concepts to be efficiently connected to the network. The third type of link weight in our network are those connecting sounds to concepts. These weights are learned by attempting to match the output of the hybrid network to semantic descriptions provided by a user community as outlined in Section 5.

We evaluate the performance of the hybrid network on a variety of information retrieval tasks for two environmental sound databases. The first database contains environmental sounds without postprocessing, where all sounds were independently described multiple times by a nonexpert user community. This allows for greater resolution in associating concepts to sounds as opposed to binary (yes/no) associations. This type of community information is what we hope to represent in the hybrid network, but collecting this data remains an arduous process and limits the size of the database.

In order to test our approach on a larger dataset, the second database consists of several thousand sound files from the *Freesound* project [16]. While this dataset is larger in terms of the numbers of sound files and semantic tags it is not as rich in terms of semantic information as tags are applied to sounds in a binary fashion by the user community. Given the noisy nature (recording/encoding quality, various levels of post production, inconsistent text labeling) of user-contributed environmental sounds, the results presented in Section 6 demonstrate that the hybrid network approach provides accurate retrieval performance. We also test performance using semantic tags that are not previously included in the network, that is, *out-of-vocabulary* tags are used as queries in text-based retrieval and as the automatic descriptions provided during annotation. Finally, conclusions and discussions of possible topics of future work are provided in Section 7.

2. An Ontological Framework Connecting Semantics and Sound

In content-based QBE, a sound query q_s is used to search a database of N sounds $\mathcal{S} = \{s_1, \ldots, s_N\}$ using a score function $F(q_s, s_i) \in \mathbb{R}$. The score function must be designed in such a way that two sound files can be compared in terms of their acoustic content. Furthermore, let $\mathcal{A}(q_s) \subset \mathcal{S}$ denote the subset of database sounds that are *relevant* to the query, while

the remaining sounds $\overline{\mathcal{A}}(q_s) \subset \mathcal{S}$ are irrelevant. In an optimal retrieval system, the score function will be such that

$$F(q_s, s_i) > F\left(q_s, s_j\right) \quad s_i \in \mathcal{A}(q_s),\ s_j \in \overline{\mathcal{A}}(q_s). \tag{1}$$

That is, the score function should be highest for sounds relevant to the query.

In text-based retrieval, the user inputs a semantic concept (descriptive tag) query q_c and the database sound set $\mathcal{S}$ is ranked in terms of relevance to the query concept. In this case, the score function $G(q_c, s_i) \in \mathbb{R}$ must relate concepts to sounds and should be designed such that

$$G(q_c, s_i) > G\left(q_c, s_j\right) \quad s_i \in \mathcal{A}(q_c),\ s_j \in \overline{\mathcal{A}}(q_c). \tag{2}$$

Once a function $G(q_c, s_i)$ is known, it can be used for the related problem of annotating unlabeled sound files. Formally, a sound query q_s is annotated using tags from a vocabulary of semantic concepts $\mathcal{C} = \{c_1, \ldots, c_M\}$. Letting $\mathcal{B}(q_s) \subset \mathcal{C}$ be the subset of concepts relevant to the query, and $\overline{\mathcal{B}}(q_s) \subset \mathcal{C}$ the irrelevant concepts, the optimal annotation system is

$$G(c_i, q_s) > G\left(c_j, q_s\right) \quad c_i \in \mathcal{B}(q_s),\ c_j \in \overline{\mathcal{B}}(q_s). \tag{3}$$

To determine effective score functions we must define the *precision* and *recall* criteria [17]. *Precision* is the number of desired sounds retrieved divided by the number of retrieved sounds and *recall* is the number of desired sounds retrieved divided by the total number of desired sounds. If we assume only one relevant object (either sound or tag) exists in the database (denoted by o_{i^*}) and the retrieval system returns only the top result for a given query, it should be clear that the probability of simultaneously maximizing precision and recall reduces to the probability of retrieving the relevant document. An optimal retrieval system should maximize this probability, which is equivalent to maximizing the posterior $P(o_i \mid q)$, that is, the relevant object is retrieved from the *maximum a posteriori* criterion, that is,

$$i^* = \underset{i \in 1:M}{\operatorname{argmax}} P(o_i \mid q). \tag{4}$$

If there are multiple relevant objects in the database, and the retrieval system returns the top R objects, we can return the objects with the R greatest posterior probabilities given the query. Thus, each of the score functions in (1)–(3) for QBE, text-based retrieval, and annotation, respectively, reduces to the appropriate posterior:

$$\begin{aligned} F(q_s, s_i) &= P(s_i \mid q_s), \\ G(q_c, s_i) &= P(s_i \mid q_c), \\ G(c_i, q_s) &= P(c_i \mid q_s). \end{aligned} \tag{5}$$

Our goal with the ontological framework proposed in this paper is to estimate all posterior probabilities of (5) in a unified fashion. This is achieved by constructing a hybrid (content/semantic) network from all elements in the sound database, the associated semantic tags, and the query (either concept or sound file) as shown in Figures 1(a)–1(c). In Figure 1(a) an audio sample is used to query the system and the output is a probability distribution over all sounds in the database. In Figure 1(b) a word is the query with the system output again a probability distribution over sounds. In Figure 1(c) a sound query is used to output a distribution over words.

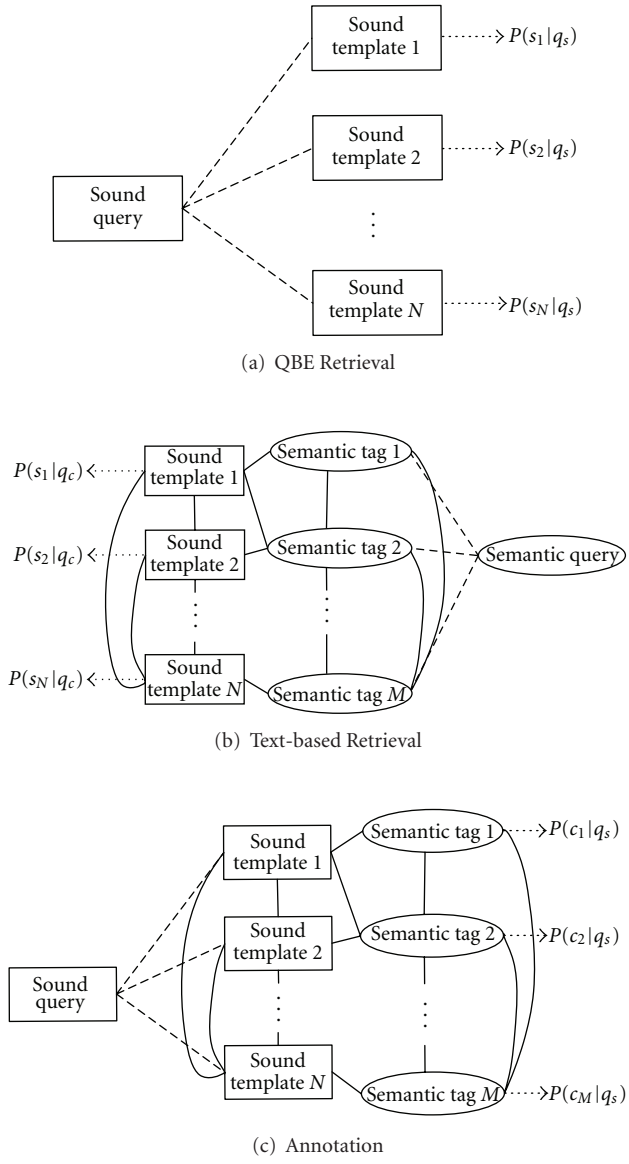


Figure 1: Operating modes of hybrid network for audio retrieval and annotation. Dashed lines indicate links added at query time, and arrows point to probabilities output by the hybrid network.

Formally, we define the hybrid network as a graph consisting of a set of nodes or vertices (ovals and rectangles in Figure 1) denoted by $\mathcal{N} = \mathcal{S} \cup \mathcal{C}$. Two nodes $i, j \in \mathcal{N}$ can be connected by an undirected link with an associated nonnegative weight (also known as length or cost), which we denote by $W(i, j) = W(j, i) \in \mathbb{R}_+$. The smaller the value of $W(i, j)$ the stronger the connection between nodes i and

j. In Figures 1(a)–1(c) the presence of an edge connecting node i to node j indicates a value of $0 \leq W(i,j) < \infty$, although the exact values for $W(i,j)$ are not indicated, while the dashed edges connecting the query node to the rest of the network are added at query time. If the text or sound file query is already in the database, then the query node will be connected through the node representing it in the network by a single link of weight zero (meaning equivalence).

The posterior distributions (score functions) from (5) are obtained from the hybrid network as

$$P(s_i \mid q_s) = \frac{e^{-d(q_s,s_i)}}{\sum_{s_j \in \mathcal{S}} e^{-d(q_s,s_j)}}, \tag{6}$$

$$P(s_i \mid q_c) = \frac{e^{-d(q_c,s_i)}}{\sum_{s_j \in \mathcal{S}} e^{-d(q_c,s_j)}}, \tag{7}$$

$$P(c_i \mid q_s) = \frac{e^{-d(q_s,c_i)}}{\sum_{c_j \in \mathcal{C}} e^{-d(q_s,c_j)}}, \tag{8}$$

where (6) is the distribution over sounds illustrated in Figure 1(a), (7) is the distribution over sounds illustrated in Figure 1(b), and (8) is the distribution over concepts illustrated in Figure 1(c). In (6)–(8), $d(q,n)$ is the path distance between nodes q and n. (Here a *path* is a connected sequence of nodes in which no node is visited more than once.) Currently, we represent $d(q,n)$ by the shortest path distance

$$d(q,n) = \min_k d_k(q,n), \tag{9}$$

where k is the index among possible paths between nodes q and n. Given starting node q, we can efficiently compute (9) for all $n \in \mathcal{N}$ using Dijkstra's algorithm [18], although for QBE (Figure 1(a)) the shortest path distance is simply the acoustic content similarity between the sound query and the template used to represent each database sound. We now describe how the link weights connecting sounds and words are determined.

3. Acoustic Information: Sound-Sound Links

As shown in Figures 1(a)–1(c), each sound file in the database is represented as a template, and the construction of these templates will be detailed in this section. Methods for ranking sound files based on the similarity of their acoustic content typically begin with the problem of acoustic feature extraction. We use the six-dimensional feature set described in [2], where features are computed from either the windowed time series data, or the short-time Fourier Transform (STFT) magnitude spectrum of 40 ms Hamming windowed frames hopped every 20 ms (i.e., 50% overlapping frames). This feature set consists of *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), *harmonicity* (a probabilistic measure of whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure), and *temporal sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one second interval).

In addition to its relatively low dimensionality, this feature set is tailored to environmental sounds while not being specifically adapted to a particular class of sounds (e.g., speech). Furthermore, we have found that these features possess intuitive meaning when searching for environmental sounds, for example, crumbling newspaper should have a high transient index and birdcalls should have high harmonicity. This intuition is not present with other feature sets, for example, it is not intuitively clear how the fourth MFCC coefficient can be used to index and retrieve environmental sounds.

Let $t \in 1 : T_j$ be the frame index for a sound file of length T_j, and $\ell \in 1 : P$ be the feature index, we define $Y_t^{(j,\ell)}$ as the ℓth observed feature for sound s_j at time t. Thus, each sound file s_j can be represented as a time series of feature vectors denoted by $Y_{1:T_j}^{(j,1:P)}$. If all sound files in the database are equally likely, the maximum-a-posteriori retrieval criterion discussed in Section 2 reduces to maximum likelihood. Thus, sound-sound link weights should be determined using a likelihood-based technique. To compare environmental sounds in a likelihood-based manner, a hidden Markov model (HMM) $\lambda^{(j,\ell)}$ is estimated from the ℓth feature trajectory of sound s_j. These HMM templates encode whether the feature trajectory varies in a constant (high or low), increasing/decreasing, or more complex (up $\rightarrow$ down; down $\rightarrow$ up) fashion. All features are modeled as conditionally independent given the corresponding HMM, that is, the likelihood that the feature trajectory of sound s_j was generated by the HMM built to approximate the simple feature trends of sound s_i is

$$\begin{aligned} L\left(s_j, s_i\right) &= \log P\left(Y_{1:T_j}^{(j,1:P)} \mid \lambda^{(i,1:P)}\right) \\ &= \sum_{\ell=1}^{P} \log P\left(Y_{1:T_j}^{(j,\ell)} \mid \lambda^{(i,\ell)}\right). \end{aligned} \tag{10}$$

Details on the estimation of $\lambda^{(i,\ell)}$ and computation of (10) are described in [2]. To make fast comparisons in the present work we allow only constant HMM templates, so $\lambda^{(i,\ell)} = \{\mu^{(i,\ell)}, \sigma^{(i,\ell)}\}$, where $\mu^{(i,\ell)}$ and $\sigma^{(i,\ell)}$ are the sample mean and standard deviation of the ℓth feature trajectory for sound s_i. Thus,

$$P\left(Y_{1:T_j}^{(j,\ell)} \mid \lambda^{(i,\ell)}\right) = \prod_{t=1}^{T_j} \gamma\left(Y_t^{(j,\ell)}; \mu^{(i,\ell)}, \sigma^{(i,\ell)}\right), \tag{11}$$

where $\gamma(y; \mu, \sigma)$ is the univariate Gaussian pdf with mean μ and standard deviation σ evaluated at y.

The ontological framework we have defined is an undirected graph, which requires weights be *symmetric* ($W(s_i,s_j) = W(s_j,s_i)$) and *nonnegative* ($W(s_i,s_j) \geq 0$). Therefore, we cannot use the log-likelihood $L(s_i,s_j)$ as the link weight between nodes s_i and s_j because it is not

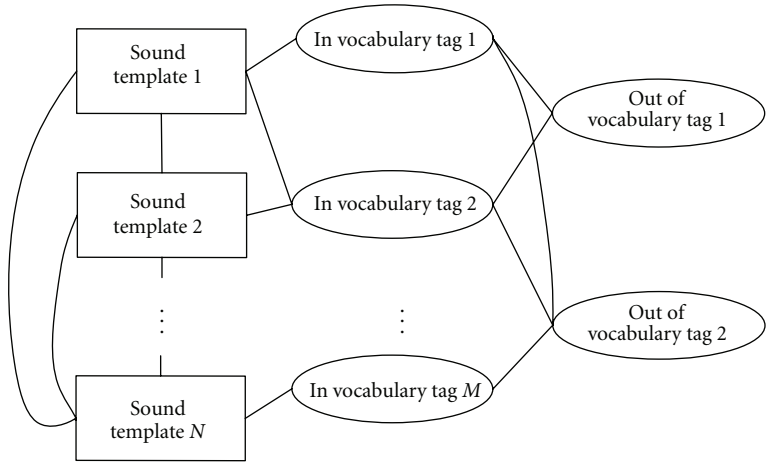


Figure 2: An example hybrid network illustrating the difference between in- and out-of-vocabulary tags.

guaranteed to be symmetric and nonnegative. Fortunately, a well-known semimetric that satisfies these properties and approximates the distance between HMM templates exists [14, 19]. Using this semimetric we define the link weight between nodes s_i and s_j as

$$W\left(s_i, s_j\right) = \frac{1}{T_i}\left[L(s_i, s_i) - L\left(s_i, s_j\right)\right] + \frac{1}{T_j}\left[L\left(s_j, s_j\right) - L\left(s_j, s_i\right)\right], \tag{12}$$

where T_i and T_j represent the length of the feature trajectories for sounds s_i and s_j, respectively. Although the semimetric in (12) does not satisfy the triangle inequality, its properties are (a) symmetry $W(s_i, s_j) = W(s_j, s_i)$, (b) nonnegativity $W(s_i, s_j) \geq 0$, and (c) distinguishability $W(s_i, s_j) = 0$ if and only if $s_i = s_j$.

4. Semantic Information: Concept-Concept Links

One technique, for determining concept-concept link weights is to a assign a link of weight zero (meaning equivalence) to concepts with common stems, for example, run/running and laugh/laughter, while other concepts are not linked. To calculate a wider variety of concept-to-concept link weights, we use a similarity metric from the WordNet::Similarity library [20]. A comparison of five similarity metrics from the WordNet::Similarity library in terms of audio information retrieval was studied in [15]. In that work the Jiang and Conrath (*jcn*) metric [21] performed best in terms of average precision, but had part of speech incompatibility problems that did not allow concept-to-concept comparisons for adverbs and adjectives. Therefore, in this work we use the *vector* metric because it supports the comparison of adjectives and adverbs, which are commonly used to describe sounds. The *vector* metric computes the cooccurrence of two concepts within the collections of words used to describe other concepts (their *glosses*) [20]. For a full review of WordNet similarity, see [20, 22].

By defining $\mathrm{Sim}(c_i, c_j)$ as the WordNet similarity between the concepts represented by nodes c_i and c_j, an appropriately scaled link weight between these nodes is

$$W\left(c_i, c_j\right) = -\log\left[\frac{\mathrm{Sim}\left(c_i, c_j\right)}{\max_{k,l}\mathrm{Sim}(c_k, c_l)}\right]. \tag{13}$$

The link weights between semantic concepts $W(c_i, c_j)$ allow the hybrid network to handle *out-of-vocabulary* tags, that is, semantic tags that were not applied to the training sound files used to construct the retrieval system can still be used either as queries in text-based retrieval or as tags applied during the annotation process. This flexibility is an important advantage of the hybrid network approach as compared to the multiclass supervised leaning approaches to audio information retrieval, for example, [7, 9]. Figure 2 displays an example hybrid network illustrating the difference between in- and out-of-vocabulary semantic tags. While out-of-vocabulary tags are connected only to in-vocabulary tags through links with weights of the form of (13), in-vocabulary tags are connected to sound files based on information from the user community via the procedure described in the following section.

5. Social Information: Sound-Concept Links

We quantify the social information connecting sounds and concepts using a $M \times N$ dimensional votes matrix V, with elements V_{ji} equal to the number of users who have tagged sound s_i with concept c_j divided by the total number of users who have tagged sound s_i. By appropriately normalizing the votes matrix, it can be interpreted probabilistically as

$$P\left(s_i, c_j\right) = \frac{V_{ji}}{\sum_k \sum_l V_{kl}}, \tag{14}$$

$$Q_{ji} = P\left(s_i \mid c_j\right) = \frac{V_{ji}}{\sum_k V_{jk}}, \tag{15}$$

$$P_{ji} = P\left(c_j \mid s_i\right) = \frac{V_{ji}}{\sum_k V_{ki}}, \tag{16}$$

where $P(s_i, c_j)$ is the joint probability between s_i and c_j, $Q_{ji} = P(s_i \mid c_j)$ is the conditional probability of sound s_i given concept c_j, and $P_{ji} = P(c_j \mid s_i)$ is defined similarly. Our goal in determining the social link weights connecting sounds and concepts $W(s_i, c_j)$ is that the hybrid network should perform both the annotation and text-based retrieval tasks in a manner consistent with the social information provided from the votes matrix. That is, the probability distribution output by the ontological framework using (7) with $q_c = c_j$ should be as close as possible to Q_{ji} from (15) and the probability distribution output using (8) with $q_s = s_i$ should be as close as possible to P_{ji} from (16). The difference between probability distributions can be computed using the Kullback-Leibler (KL) divergence.

We define $\mathbf{w} = \{W(s_i, c_j) \mid V_{ji} \neq 0\}$ to be the vector of all sound-word link weights, $\hat{Q}_{ji}(\mathbf{w})$ as the probability distribution output by the ontological framework using (7) with $q_c = c_j$, and $\hat{P}_{ji}(\mathbf{w})$ as the probability distribution output by the ontological framework using (8) with $q_s = s_i$. Treating concept s_i as the query, the KL divergence between the distribution over database sounds obtained from the network and the distribution obtained from the user votes matrix is

$$\mathrm{KL}(s_i, \mathbf{w}) = \sum_{c_j \in \mathcal{C}} P_{ji} \log \left[\frac{P_{ji}}{\hat{P}_{ji}(\mathbf{w})} \right]. \tag{17}$$

Similarly, given concept c_j as the query, the KL divergence between the distribution of concepts obtained from the network and the distribution obtained from the user votes matrix is

$$\mathrm{KL}\left(c_j, \mathbf{w}\right) = \sum_{s_i \in \mathcal{S}} Q_{ji} \log \left[\frac{Q_{ji}}{\hat{Q}_{ji}(\mathbf{w})} \right]. \tag{18}$$

The network weights are then determined by solving the optimization problem

$$\min_{\mathbf{w}} \sum_{s_i \in \mathcal{S}} \sum_{c_j \in \mathcal{C}} \mathrm{KL}(s_i, \mathbf{w}) + \mathrm{KL}\left(c_j, \mathbf{w}\right). \tag{19}$$

Empirically, we have found that setting the initial weight values to $W(s_i, c_j) = -\log P(s_i, c_j)$, leads to quick convergence. Furthermore, if resources are not available to use the KL weight learning technique, setting the sound-concept link weights to $W(s_i, c_j) = -\log P(s_i, c_j)$ provides a simple and effective approximation of the optimized weight.

Presently, the votes matrix is obtained using only a simple tagging process. In the future we hope to augment the votes matrix with other types of community activity, such as discussions, rankings, or page navigation paths on a website. Furthermore, sound-to-concept link weights can be set as design parameters rather than learned from a "training set" of tags provided by users. For example, expert users can make sounds equivalent to certain concepts through the addition of zero-weight connections between specified sounds and concepts, thus, improving query results for nonexpert users.

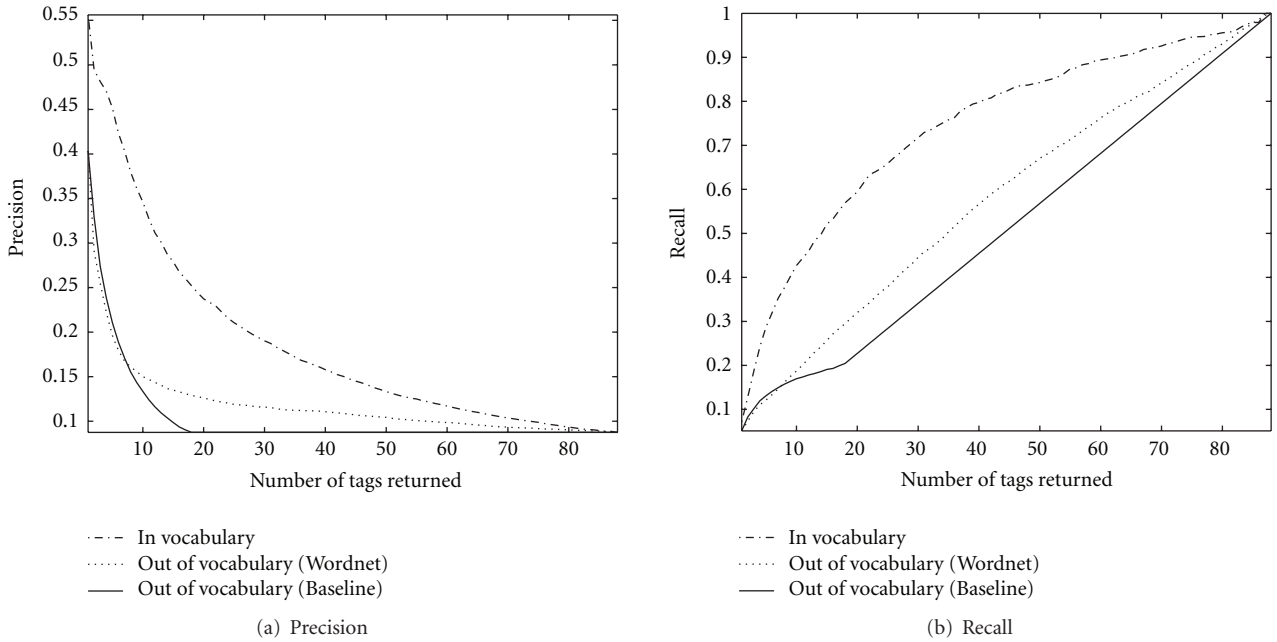
6. Results and Discussion

In this section, the performance of the hybrid network on the annotation and text-based retrieval tasks will be evaluated. (QBE results were considered in our previous work [2] and are not presented here).

6.1. Experimental Setup. Two datasets are used in the evaluation process. The first dataset, which we will refer to as the *Soundwalks* data set contains 178 sound files uploaded by the authors to the *Soundwalks.org* website. The 178 sound files were recorded during seven separate field recording sessions, lasting anywhere from 10 to 30 minutes each and sampled at 44.1 KHz. Each session was recorded continuously and then hand-segmented by the authors into segments lasting between 2–60 s. The recordings took place at three light rail stops (75 segments), outside a stadium during a football game (60 segments), at a skatepark (16 segments), and at a college campus (27 segments). To obtain tags, study participants were directed to a website containing ten random sounds from the set and were asked to provide one or more single-word descriptive tags for each sound. With 90 responses, each sound was tagged an average of 4.62 times. We have used 88 of the most popular tags as our vocabulary.

Because the Soundwalks dataset contains 90 subject responses, a nonbinary votes matrix can be used to determine the sound-concept link weights described in Section 5. Obtaining this votes matrix requires large amounts of subject time, thus, limiting its size. To test the hybrid network performance on a larger dataset, we use 2064 sound files and a 377 tag vocabulary from *Freesound.org*. In the Freesound dataset tags are applied in a binary (yes/no) manner to each sound file by users of the website. The sound files were randomly selected from among all files (whether encoded in a lossless or lossy format) on the site containing any of the 50 most used tags and between 3–60 seconds in length. Additionally, each sound file contained between three and eight tags, and each of the 377 tags in the vocabulary were applied to at least five sound files.

To evaluate the performance of the hybrid network we adopt a two-fold cross validation approach where all of the sound files in our dataset are partitioned into two nonoverlapping subsets. One of these subsets and its associated tags is then used to build the hybrid network via the procedure described in Sections 2–5. The remaining subset is then used to test both the annotation and text-based retrieval performance for unlabeled environmental sounds. Furthermore, an important novelty in this work is the ability of the hybrid network to handle *out-of-vocabulary* tags. To test performance for out-of-vocabulary tags, a second tier of cross validation is employed where all tags in the vocabulary are partitioned into five random, nonoverlapping subsets. One of these subsets is then used along with the subset of sound files to build the hybrid network, while the remaining tags are held out of vocabulary. This partitioning procedure is summarized in Table 1 for both the Soundwalks and Freesound datasets. Reported results are the average over these 10 (five tag, two sound splits) cross-validation runs.



(a) Precision (b) Recall

Figure 3: Precision and recall curves for annotation of unlabeled sound files in the *Soundwalks* dataset averaged over 10 cross-validation splits.

Table 1: Database partitioning procedure for each cross validation run.

	Soundwalks	Freesound
Number of sound files	178	2064
In network (training)	89	1032
Out of network (testing)	89	1032
Number of tags	88	377
In vocabulary	18	75
Out of vocabulary	70	302

Relevance is determined to be positive if a held out sound file was actually labeled with a tag. It is also important to note that the tags for both datasets are not necessarily provided by expert users, thus, our relevance data can be considered "noisy."

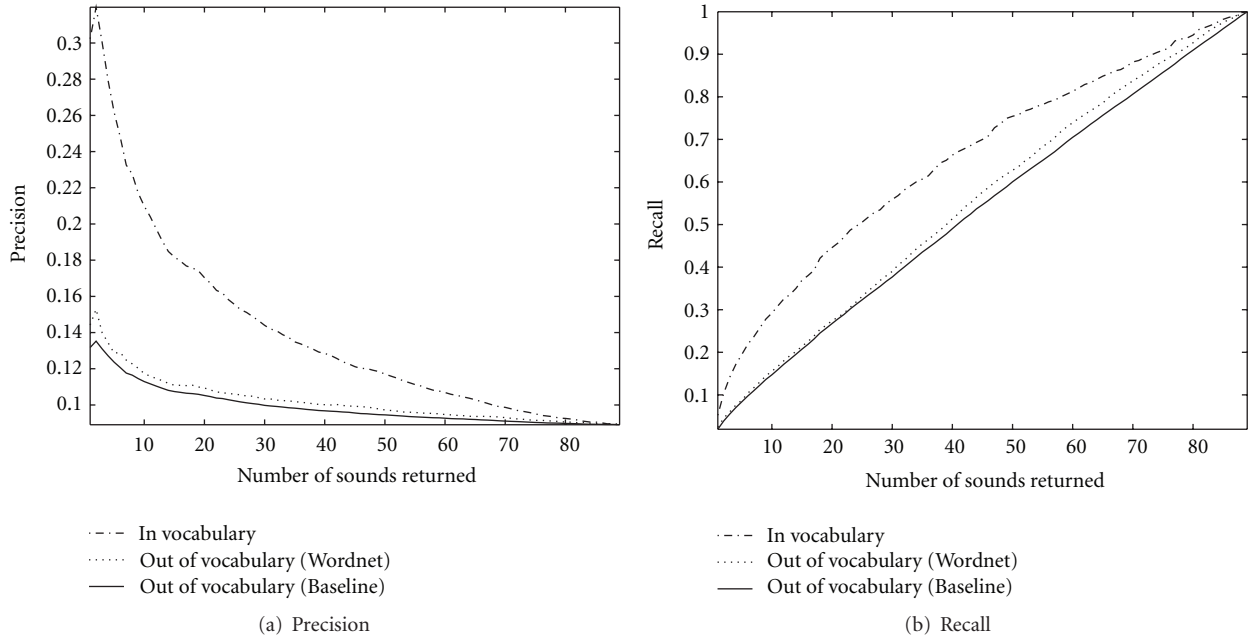
6.2. Annotation. In annotation each sound in the testing set is used as a query to provide an output distribution over semantic concepts. For a given sound query q_s we denote by $\mathcal{B}(q_s)$ the set of tags, and $|\mathcal{B}|$ the number of relevant tags for that query. Assuming M tags in a database are ranked in order of decreasing probability for a given query, by truncating the list to the top n tags, and counting the number of relevant tags, denoted by $|\mathcal{B}^{(n)}|$, we define $precision = |\mathcal{B}^{(n)}|/n$ and $recall = |\mathcal{B}^{(n)}|/|\mathcal{B}|$. Average precision is found by incrementing n and averaging the precision at all points in the ranked list where a relevant sound is located. Additionally, the area under the receiver operating characteristics curve (AROC) is found by integrating the ROC curve, which plots the true positive versus false positive rate for the ranked list of output tags.

Figures 3(a) and 3(b) display the precision and recall curves, respectively, averaged over all sound queries and cross-validation runs for the soundwalks dataset. The three curves in Figure 3 represent three different ways of building the hybrid network. The *in-vocabulary* curve can be considered as an upper bound of annotation performance as all tags are used in building the network. The *out-of-vocabulary (WordNet)* curve uses only a subset of tags to build the hybrid network, and remaining tags are connected only through concept-concept links as described in Section 4. The *out-of-vocabulary (Baseline)* curve uses only a subset of tags to build the hybrid network, and remaining tags are returned in random order. This is how the approach of training a classifier for each tag, for example, [7–9] would behave for out of vocabulary tags. From Figures 3(a) and 3(b) we see that out-of-vocabulary performance is improved both in terms of precision and recall when WordNet link weights are included. Additionally, from the precision curve of Figure 3(a) we see that approximately 15% of the top 20 out of vocabulary tags are relevant, while for in vocabulary tags this number is 25%. Considering the difficulty of the out of vocabulary problem, and that each sound file is labeled with much less than 20 tags this performance is quite promising. From the recall curve of Figure 3(b) approximately 30% of relevant out-of-vocabulary tags are returned in the top 20, compared to approximately 60% of in-vocabulary tags.

Table 2 contains the mean average precision (MAP) and mean area under the receiver operating characteristics curve (MAROC) values for both the Soundwalks and Freesound databases. We see that performance is comparable between the two datasets, despite the Freesound set being an order

Table 2: Annotation performance using out-of-vocabulary semantic concepts.

	Soundwalks		*Freesound*	
	MAP	MAROC	MAP	MAROC
In vocabulary (upper bound)	0.4333	0.7523	0.4113	0.8422
Out of vocabulary (WordNet)	0.2131	0.6322	0.1123	0.6279
Out of vocabulary (Baseline)	0.1789	0.5353	0.1092	0.5387



(a) Precision (b) Recall

Figure 4: Recall and precision curves for text-based retrieval of unlabeled sound files in the *Soundwalks* dataset averaged over 10 cross-validation splits.

of magnitude larger. The slightly better performance on the Soundwalks dataset is most likely due to the large amount of social information contained in the votes matrix, which is used to set sound-concept link weight values. The in-vocabulary MAP values of 0.4333 and 0.4113 compare favorably to the per-word MAP value of 0.179 reported in [7] for annotating BBC sound effects. Benchmarking the performance for out-of-vocabulary tags is more difficult since this task is often not considered in the literature.

6.3. Text-Based Retrieval. In text-based retrieval each semantic tag is used as a query to provide an output distribution over the test sounds. For a given query we denote by $\mathcal{A}(q_c)$ the set of relevant test sounds that are labeled with the query word, and $|\mathcal{A}|$ as the number of relevant test sounds for that query. Precision, recall, MAP, and MAROC values are then computed as described above. Figures 4(a) and 4(b) display the precision and recall curves, respectively, averaged over all sound queries and cross-validation runs for the Soundwalks dataset, while Table 3 displays the MAP and MAROC values. As with annotation, text-based retrieval with out-of-vocabulary concepts is significantly more difficult than with in vocabulary concepts, but including the concept-concept links based on the measure of WordNet similarity helps to ameliorate retrieval performance.

To demonstrate that retrieval performance is most likely considerably better than the reported precision, recall, MAP, and MAROC performance averaged over noisy tags contributed by nonexpert users, we provide the example of Table 4. Here, the word "rail" is used as an out-of-vocabulary query to retrieve unlabeled sounds, and the top four results are displayed. Additionally, Table 4 displays the posterior probability of each of the top four results, the shortest path of nodes from the query to the output sounds, and whether or not the output sound is relevant. The top result is the sound mixture of automobile traffic and a train horn, but is not tagged by any users with the word "rail," even though like the sounds actually tagged with "rail" it is a recording of a train station. Although filtering these types of results would improve quantitative performance it would require listening to thousands of sound files and overruling subjective decisions made by the users who listened to and labeled to the sounds.

6.4. In-Vocabulary Semantic Information. Effective annotation and retrieval for out-of-vocabulary tags requires some

Table 3: Text-based retrieval performance using out-of-vocabulary semantic concepts.

	Soundwalks		*Freesound*	
	MAP	MAROC	MAP	MAROC
In vocabulary (upper bound)	0.2725	0.6846	0.2198	0.7100
Out of vocabulary (WordNet)	0.1707	0.6291	0.0681	0.5788
Out of vocabulary (Baseline)	0.1283	0.5355	0.0547	0.5414

Table 4: Top four results from *Soundwalks* data set for text-based retrieval with out of vocabulary query "rail". Parenthetical descriptions are not actual tags, but provided to give an idea of the acoustic content of the sound files.

Posterior probability	Node path	Relevant
0.19	rail⇒train⇒segment94.wav (*train bell*)⇒segment165.wav (*traffic/train horn*)	No
0.17	rail⇒voice⇒segment136.wav (*pa announcement*)⇒segment133.wav (*pa announcement*)	Yes
0.15	rail⇒train⇒segment40.wav (*train brakes*)⇒segment30.wav (*train bell/brakes*)	Yes
0.09	rail⇒train⇒segmen40.wav (*train brakes*)⇒segment147.wav (*train horn*)	Yes

Table 5: Performance of retrieval tasks with the *Soundwalks* dataset using WordNet connections between in-vocabulary semantic concepts.

	Text-based retrieval		Annotation	
	MAP	MAROC	MAP	MAROC
With WordNet	0.2166	0.6133	0.2983	0.6670
Without WordNet	0.3744	0.6656	0.4633	0.7978

method of relating the semantic similarity of tags, for example, the WordNet similarity metric used in this work. In this section we examine how the inclusion of semantic connections between in-vocabulary tags affects annotation and text-based retrieval performance. Table 5 compares the MAP and MAROC values for the Soundwalks dataset where all tags are used in building the network both with and without semantic links connecting tags. The results of Table 5 suggest that when the information connecting sounds and tags is available (i.e., tags are in the vocabulary) the semantic links provided by WordNet confound the system by allowing for possibly irrelevant relationships between tags. This is not unlike the observations of [23] where using WordNet did not significantly improve information retrieval performance. Comparing the environmental sound retrieval performance of WordNet similarity with other techniques for computing prior semantic similarity (e.g., Google distance [24]) remains a topic of future work, since some measure of semantic similarity is necessary to handle out-of-vocabulary tags.

7. Conclusions and Future Work

Currently, a significant portion of freely available environmental sound recordings are user contributed and inherently noisy in terms of audio content and semantic descriptions. To aid in the navigation of these audio databases we show the utility of a system that can be used for text-based retrieval of unlabeled audio, content-based query-by-example, and automatic audio annotation. Specifically, an ontological framework connects sounds to each other based on a measure of perceptual similarity, tags are linked based on a measure of semantic similarity, while tags and sounds are connected by optimizing link weights given user preference data. An advantage of this approach is the ability of the system to flexibly extend when new sounds and/or tags are added to the database. Specifically, unlabeled sound files can be queried or annotated with out-of-vocabulary concepts, that is, tags that do not currently exist in the database.

One possible improvement to the hybrid network structure connecting semantics and sound might be achieved by exploring different link weight learning techniques. Currently, we use a "divide and conquer" approach where the three types of weights (sound-sound, concept-concept, sound-concept) are learned independently. This could lead to scaling issues, especially if the network is expanded to contain different node types. One possible approach to overcome these scaling issues could be to learn a dissimilarity function from ranking data [25]. For example, using the sound similarity, user preference, and WordNet similarity data to find only rankings between words and sounds of the form "A is more like B than C is D", we can learn a single dissimilarity function for the entire network that preserves this rank information.

Another enhancement would be to augment the hybrid network with a recursive clustering scheme such as those described in [26]. We have successfully tested this approach in [14], where each cluster becomes a node in the hybrid network, and all sounds assigned to each cluster are connected to the appropriate cluster node by a link of weight zero. These cluster nodes are then linked to the nodes

representing semantic tags. While this approach limits the number of sound-tag weights that need to be learned, the additional cluster nodes and links tend to cancel out this savings. Furthermore, when a new sound is added to the network we still must compute its similarity to all sounds previously in the network (this is also true for new tags). For sounds, it might be possible to represent each sound file and sound cluster as a Gaussian distribution, and then use symmetric Kullback-Leibler divergence to calculate the link weights connecting new sounds added to the network to preexisting clusters. Unfortunately, this approach would not extend to the concept nodes in the hybrid network as we currently know of no technique for representing a semantic tag as a Gaussian, even though the WordNet similarity metric could be used to cluster the tags. Perhaps a technique where a fixed number of sound/tag nodes are sampled to have link weights computed each time a new sound/tag is added to the network could help make the ontological framework more computationally efficient. A link weight pruning approach might also help improve computational complexity.

Finally, using a domain-specific ontology such as the MX music ontology [27] might be better suited to audio information retrieval than a purely lexical database such as WordNet. For environmental sounds, the theory of soundscapes [28, 29] might be a convenient first step, as the retrieval system could be specially adapted to the different elements of a soundscape. For example, sounds such as traffic and rain could be connected to a *keynote* sublayer in the hybrid network, while sounds such as alarms and bells could be connected to the *sound signal* sublayer. Once the subjective classification of sound files into the different soundscape elements are obtained adding this sublayer into the present ontological framework could be an important enhancement to the current system.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants NSF IGERT DGE-05-04647 and NSF CISE Research Infrastructure 04-03428. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

References

[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, Article ID 4472077, pp. 668–696, 2008.

[2] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 688–707, 2010.

[3] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet, "A game-based approach for collecting semantic annotations of music," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR '07)*, Vienna, Austria, 2007.

[4] M. I. Mandel and D. P. W. Ellis, "A Web-based game for collecting music metadata," *Journal of New Music Research*, vol. 37, no. 2, pp. 151–165, 2008.

[5] M. Slaney, "Semantic-audio retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 4, pp. 4108–4111, Orlando, Fla, USA, 2002.

[6] B. Whitman and D. Ellis, "Automatic record reviews," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR '04)*, pp. 470–477, 2004.

[7] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, Article ID 4432652, pp. 467–476, 2008.

[8] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 37–40, New Paltz, NY, USA, 2009.

[9] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st International ACM Conference on Multimedia Information Retrieval (MM '08)*, pp. 105–112, Vancouver,Canada, August 2008.

[10] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, P. Herrera, and N. Wack, "Nearest-neighbor generic sound classification with a WordNet-based taxonomy," in *Proceedings of the 116th AES Convention*, Berlin, Germany, 2004.

[11] E. Martinez, O. Celma, M. Sordo, B. de Jong, and X. Serra, "Extending the folksonomies of freesound.org using content-based audio analysis," in *Proceedings of the Sound and Music Computing Conference*, Porto, Portugal, 2009.

[12] WordNet, http://wordnet.princeton.edu/.

[13] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Mass, USA, 1998.

[14] G. Wichern, H. Thornburg, and A. Spanias, "Unifying semantic and content-based approaches for retrieval of environmental sounds," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, pp. 13–16, New Paltz, NY, USA, 2009.

[15] B. Mechtley, G. Wichern, H. Thornburg, and A. S. Spanias, "Combining semantic, social, and acoustic similarity for retrieval of environmental sounds," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10)*, 2010.

[16] Freesound, http://www.freesound.org/.

[17] C. J. V. Rijsbergen, *Information Retrieval*, Butterwoths, London, UK, 1979.

[18] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press and McGraw-Hill, Cambridge, UK, 2nd edition, 2001.

[19] B. H. Huang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Technical Journal*, vol. 64, no. 2, pp. 1251–1270, 1985.

[20] T. Pederson, S. Patwardhan, and J. Michelizzi, "Wordnet:similarity—measuring the relatedness of concepts," in *Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference (IAAI '04)*, pp. 1024–1025, AAAI Press, Cambridge, MA, USA, 2004.

[21] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X '97)*, pp. 19–33, Taiwan, 1997.

[22] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures," in *Proceedings of the Workshop on WordNet and Other Lexical Resources, 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittburgh, Pa, USA, 2001.

[23] R. Mandala, T. Tokunaga, and H. Tanaka, "The use of wordnet in information retrieval," in *Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems*, pp. 31–37, Montreal, Canada, 1998.

[24] R. L. Cilibrasi and P. M. B. Vitányi, "The Google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[25] H. Ouyang and A. Gray, "Learning dissimilarities by ranking: from SDP to QP," in *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 728–735, Helsinki, Finland, July 2008.

[26] J. Xue, G. Wichern, H. Thornburg, and A. Spanias, "Fast query by example of environmental sounds via robust and efficient cluster-based indexing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 5–8, Las Vegas, Nev, USA, April 2008.

[27] A. Ferrara, L. A. Ludovico, S. Montanelli, S. Castano, and G. Haus, "A semantic web ontology for context-based classification and retrieval of music resources," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 3, pp. 177–198, 2006.

[28] R. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*, Destiny Books, Rochester, Vt, USA, 1994.

[29] B. Truax, *Acoustic Communication*, Ablex Publishing, Norwood, NJ, USA, 1984.