# SHORTEST PATH TECHNIQUES FOR ANNOTATION AND RETRIEVAL OF ENVIRONMENTAL SOUNDS

**Brandon Hawley**
Arizona State University
Computer Science (SCIDSE)
bmechtley@asu.edu

**Perry Cook**
Princeton University
Computer Science and Music
prc@cs.princeton.edu

**Andreas Spanias**
Arizona State University
Electrical Engineering (SECEE)
spanias@asu.edu

## ABSTRACT

Many techniques for text-based retrieval and automatic annotation of music and sound effects rely on learning with explicit generalization, training individual classifiers for each tag. Non-parametric approaches, where queries are individually compared to training instances, can provide added flexibility, both in terms of robustness to shifts in database content and support for foreign queries, such as concepts not yet included in the database. In this paper, we build upon prior work in designing an ontological framework for annotation and retrieval of environmental sounds, where shortest paths are used to navigate a network containing edges that represent content-based similarity, semantic similarity, and user tagging data. We evaluate novel techniques for ordering query results using weights of both shortest paths and minimum cost paths of specified lengths, pruning outbound edges by nodes' K nearest neighbors, and adjusting edge weights depending on type (acoustic, semantic, or user tagging). We evaluate these methods both through traditional cross-validation and through simulation of live systems containing a complete collection of sounds and tags but incomplete tagging data.

## 1. INTRODUCTION

### 1.1 Multiclass and non-parametric retrieval

Many techniques for text-based retrieval or classification of audio signals are parametric in nature, relying on explicit generalization, where individual classifiers are created for each label. For example, classification systems have been built for automatic record reviews [18], onomatopoetic labels [9], and genre [17], emotion [10], and instrumentation [5,7] identification. These systems make use of techniques such as one-versus-all discrimination [18], training each label with a support vector machine (SVM) classifier [1, 9], and learning a separate gaussian mixture model (GMM) for each label [16, 18].

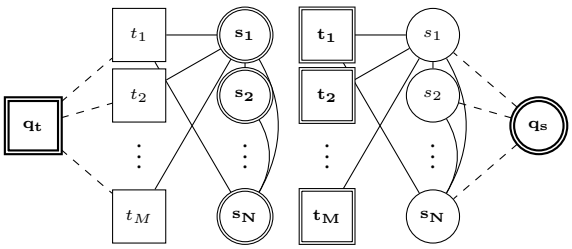
These multiclass methods benefit from constant query time complexity independent of the number of training instances, in that it is only necessary for each query (such as a sound, in the case of annotation) to be measured against each classifier. For specific, relatively stationary label domains, such as musical genres, this can be seen as a great benefit, especially when the number of sounds greatly exceeds the number of labels. However, there are many cases where non-parametric models can provide additional flexibility and robustness. One such case involves the presence of multiple types of information beyond acoustic feature vectors and annotations. For example, [19] and [13] describe methods where similarity between semantic concepts can assist in retrieval and annotation using tags not yet seen in training data. In large-scale systems with more complete tag sets, this may be less of a problem, but in live databases with incomplete tagging where no large-scale training database exists beyond user activity, as in the case of Freesound [1] , retrieval results can often come up empty.

Non-parametric (also known as similarity- or instance-based [6, 11]) schemes compare each query to instances in a live database rather than having distinct training and production / evaluation stages. For example, [2], [3], and [12] use K-nearest-neighbors retrieval, where unlabeled sounds are annotated with tags belonging to their nearest neighbors in an acoustic feature space. [15] and [2] build two separate hierarchical cluster models—one for retrieval and one for annotation.

### 1.2 Associative retrieval

Graph-based techniques are often used for search in semantic and other associative networks. One technique that has seen much use is spreading activation. In spreading activation, an initial node (a query) is labeled with some weight, and this weight is spread to neighboring nodes with some decay. Spreading activation has been used in information retrieval applications, where nodes correspond to documents and terms [4]. Shortest paths are also of interest in associative retrieval. [20] introduces a graph-based framework where sounds are connected to tags through user activity and sounds are fully connected via acoustic similarity estimated by an HMM-based query-by-example algorithm described in [21]. New queries are immediately connected to other sounds or tags (either through acoustic or semantic similarity via the WordNet::Similarity library [14]), and shortest path distances using all nodes are

[1] Freesound: http://freesound.org/




(a) Text-based retrieval (b) Annotation

**Figure 1**: Two different possible query tasks with a single retrieval network. $q_s$ and $q_t$ represent a sound query and a tag query, respectively, and $t_1, t_2, ..., t_M \in T$ and $s_1, s_2, ..., s_N \in S$ represent tag and sound nodes already in the database. The query node and the subset of nodes over which the user is querying is marked in bold, and on-demand edge weights between the query and its respective class of nodes (sounds or tags) are marked with dashed lines. Note that $S$ forms a clique.

used to rank retrieval results.

The results of [19] demonstrate that including semantic similarity to account for tags foreign to the training set can assist in annotation and text-based retrieval both for a random subset from the Freesound library, where tags are associated to sounds in a binary manner, and a smaller comprehensive user study, where each sound is tagged by multiple users. In this paper, we build upon this graph-based technique and perform a more in-depth study of its properties. Namely, we seek to a) evaluate the system using both traditional cross-validation and simulations of real-world systems with complete sound and tag sets but incomplete tagging data, b) demonstrate the effectiveness of adding a shortest-path algorithm to any existing tag-based query system, regardless of the presence of acoustic and semantic similarity measures, c) improve shortest-path retrieval performance by pruning network edges to nodes' K nearest neighbors, and d) explore the impact of assigning different weights to the importance of acoustic, semantic, and user-provided information.

## 2. SHORTEST PATH RETRIEVAL

### 2.1 Network structure

Formally, the network structure for retrieval and annotation takes the form of a weighted, undirected graph, $G = (V, E)$, where $V = S \cup T$ and $S$ and $T$ represent sets of sound and tag nodes. The graph edges, $E$, can be partitioned into three disjoint subsets, $E_{SS} \subseteq S \times S$, $E_{ST} \subseteq S \times T$, and $E_{TT} \subseteq T \times T$, representing acoustic, user-provided, and semantic information. The weighting function is denoted by $w : E \rightarrow \mathbb{R}^+$. This type of network structure can be adapted to different domains, such as music or even text documents, but for the sake of this paper, we focus on the task of retrieving and annotating environmental sounds. We therefore assume that sounds take the form of short audio clips representing individual sonic events. For more discussion on the concept of a "sound event," see the related discussion in [21]. In the following sections, we will discuss how the weights for $E_{SS}$, $E_{ST}$, and $E_{TT}$ are calculated.

#### *2.1.1 Sound-to-sound weights ($E_{SS}$)*

Sound-to-sound weights can be computed by comparing the acoustic content of each sound. For this task, we use the Sirens library [2] . For detailed information of how sound similarity is computed in Sirens and an evaluation of its performance, see [21]. A summary is as follows:

Sirens begins with acoustic feature extraction, where six features are calculated on overlapping 40ms Hamming windows hopped every 20ms. The feature trajectory for a sound file is given by $Y_{1:T}^{(1:F)}$, where $Y_t^{(i)}$ is the $i$-th feature's value at frame $t$.

The six features used by Sirens include *loudness*, the dB-scaled RMS level over time; *temporal sparsity*, the ratio of $l^\infty$ and and $l^1$ norms calculated over all windowed RMS levels in a one-second interval; *spectral sparsity*, the ratio of $l^\infty$ and $l^1$ norms calculated over short-time Fourier transform (STFT) magnitudes; *spectral centroid*, the bark-weighted average spectral content of a sound at any point in time; *transient index*, the $l^2$ norm of the difference of Mel frequency cepstral coefficients (MFCC) between consecutive frames; and *harmonicity*, a probabilistic measure of whether or not the signal comes from a harmonic source. This feature set was developed for a broad range of environmental sounds rather than any specific class of sounds and with a focus on ecological validity such that the features would best relate to qualities of human audition by using perceptual scalings of spectral content [21].

To compare sounds, [21] describes a method of estimating $L(s_i, s_j) = -\log P(Y_{1:T}^{(1:F)}(s_i)|\lambda^{(1:F)}(s_j))$, the log-likelihood of the feature trajectory of sound $s_i$ being generated from the hidden Markov model (HMM), $\lambda^{(1:F)}(s_j)$, built to approximate the simple (i.e. constant, linear, or quadratic) feature trends of sound $s_j$. For retrieval in our undirected graph, however, it is helpful to have a semi-metric between sounds that is symmetric and nonnegative. In [8], a semi-metric that holds these properties is given:

$$w(s_i, s_j) = L(s_i, s_i) + L(s_j, s_j) - L(s_i, s_j) - L(s_j, s_i). \quad (1)$$

#### *2.1.2 Sound-to-tag weights ($E_{ST}$)*

Letting $U_{|S|\times|T|}$ be a votes matrix where $U_{ij}$ is equal to the number of users who have tagged sound $s_i$ with tag $t_j$, we can compute the joint probability of $s_i$ and $t_j$ as

$$P(s_i, t_j) = \frac{U_{ij}}{\sum_{k,l} U_{kl}} \quad (2)$$

$$w(s_i, t_j) = -\log P(s_i, t_j). \quad (3)$$

#### *2.1.3 Tag-to-tag weights ($E_{TT}$)*

The results from [13] and [19] demonstrate that semantic similarity (tag-to-tag edges) obtained from WordNet::-

[2] Sirens (Segmentation, Indexing, and Retrieval of Environmental Sounds): http://github.com/plant/sirens

Similarity [14] scores can be useful when performing text-based retrieval queries using tags not in the database or annotating sounds with these foreign tags. In general, however, it was found that including tag-to-tag links between in-network tags can hinder performance, so we have excluded these links in this paper, as we are chiefly interested in studying the effects of different shortest-path retrieval strategies rather than the source of the weights themselves.

### 2.2 Shortest path retrieval

Given the structure of the graph, $G(V, E)$ and its weights, $w$, as defined in the previous section, we rank search results according to their shortest path lengths from the query, $q$, to the target, $t$, in ascending order:

$$w^*(q,t) = \min_{P=\langle q,\ldots,t\rangle} \sum_{i=1}^{|P|-1} w(P_i, P_{i+1}), \qquad (4)$$

## 3. NETWORK MODIFICATIONS

### 3.1 Depth-ordered retrieval

Shortest paths may sometimes hinder retrieval results in cases where they provide discursive paths that rely on numerous relations. For example, if sound-to-tag weights are trained with ground truth data obtained from user studies or extensive user activity, it would be desirable to only use these direct paths and visit no other nodes rather than second-guessing users (who, for the purpose of evaluation, we often assume are experts).

In these cases, we can form a list, $L$, of positive integers representing desired path depths, with an optional final element, $*$, representing shortest paths of any depth. Any targets unconnected to the query will be returned at the end of the list in random order. For example, $L = (2, *)$ will prioritize minimum-cost direct edges between the query and targets first, only using shortest paths as a last resort in the absence of direct edges. $L = (2, 3, *)$ will first return all direct edges, then shortest paths containing only three nodes, and finally all shortest paths. For the case of $L = (2, 3, ...)$, where depths are in monotonically increasing order, this algorithm performs similarly to a breadth-first search. In Section 4.3, we will discuss the relative performance of depth orderings.

### 3.2 Edge pruning

Shortest-path retrieval can be quite computationally expensive. In the worst case, Dijkstra's algorithm has $O(|E| + |V|\log|V|)$ time complexity. As our graph is quite well-connected (for large numbers of sounds), we can assume the complexity of performing a single query is $O(|V|^2)$, as $|E| = O(|V|^2)$ when the number of sounds greatly exceeds the number of tags. In these cases, it can be beneficial to limit search to only a node's $K$ nearest neighbors, giving a complexity of $O(K|V| + |V|\log|V|) = O(|V|\log|V|)$.[3] In Section 4.4, we will discuss the effects of $K$ nearest neighbor pruning, where $G$ is converted to a directed graph with inbound and outbound edges identical to the original undirected edges and all but the $K$ lowest-weight outbound edges are removed.

[3] Using spectral clustering to cluster sounds, as in [20], we can even improve this to $O(\log|V|)$ complexity.

### 3.3 Weighting edge classes

Lastly, it should be noted that the ranking of search results can be quite sensitive to variations in weighting between the different classes of edges, $E_{SS}$, $E_{ST}$, and $E_{TT}$, as each assumes a different probabilistic model. If one class has particularly low weights, its edges may be used more frequently than edges of other classes. In Section 4.5, we examine the effects of setting class-specific weights, $\gamma_C$:

$$w_\gamma(n_1, n_2) = \gamma_C w(n_1, n_1)$$
$$\forall (n_1, n_2) \in E_C, \forall E_C \in \{E_{SS}, E_{ST}, E_{TT}\} \qquad (5)$$

## 4. RESULTS AND DISCUSSION

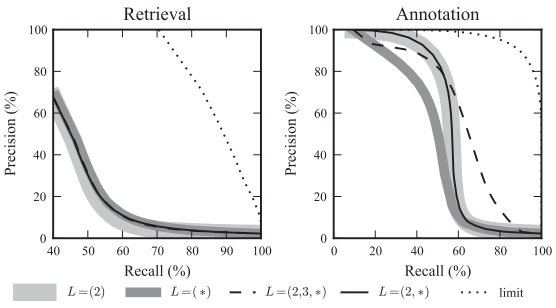
### 4.1 Training data

To evaluate the text-based retrieval and annotation performance of our various modifications to the shortest-path retrieval method, we use two datasets that link sounds to conceptual tags. The first dataset, *Soundwalks*, is a collection of 178 2-60s sound events manually segmented from one of four outdoor recording sessions recorded at 44.1KHz. Sound environments include light rail stops, a stadium during a football game, a skatepark, and a walkway on a college campus. To obtain tagging data, an online user study was performed where each user was asked to provide several semantic tags for 10 randomly selected sound events by freely typing terms separated by commas. Terms that could not be found in the WordNet taxonomy were ignored. With a total of 90 responses, each sound was tagged an average of 4.62 times. In [19] and [13], only the most popular 88 tags were used for evaluation, but we have used the entire set of 612 tags to more accurately study the system's performance. The second dataset, *Freesound*, was obtained from user activity on the website *Freesound.org*. 2046 sounds were randomly selected from the set of all sounds 2-60s in duration and containing at least one of any of the 50 most popular tags on the site. Each sound is associated with 3-8 tags, and each of the 377 total tags used is associated with at least 5 sounds. Note that on *Freesound*, tags are only associated to sounds in a binary manner, so all nonzero sound-to-tag weights are equal. Both datasets have been used to test the performance of the system a) against a slightly more reliable ground truth, in the case *Soundwalks*, where each sound file has been tagged by 4-5 users, and b) against a larger collection of sounds, as in *Freesound*.

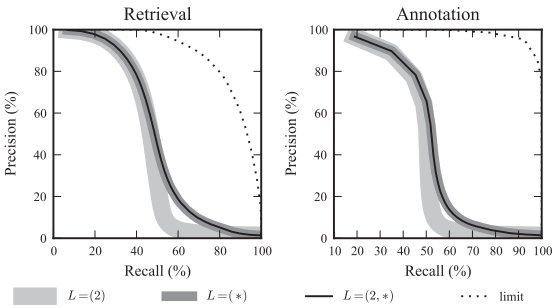
### 4.2 Evaluation methodology

#### *4.2.1 Cross-validation versus incomplete tagging data*

For multiclass retrieval, where classifiers are trained for each search term, evaluation procedures typically involve




(a) Soundwalks Precision/Recall




(b) Freesound Precision/Recall

| | Retrieval | | Annotation | |
|---|---|---|---|---|
| $L$ | MAP | MAROC | MAP | MAROC |
| $(2)$ | 0.5505 | 0.7615 | 0.5920 | 0.7814 |
| $(*)$ | **0.5751** | **0.7910** | 0.5230 | 0.7833 |
| $(2,*)$ | 0.5673 | 0.7887 | 0.5964 | 0.7919 |
| $(2,3,*)$ | 0.5678 | 0.7849 | **0.6590** | **0.8824** |

(c) Soundwalks MAP/MAROC

| | Retrieval | | Annotation | |
|---|---|---|---|---|
| $L$ | MAP | MAROC | MAP | MAROC |
| $(2)$ | 0.5200 | 0.7504 | 0.5237 | 0.7506 |
| $(*)$ | 0.5594 | 0.8610 | **0.5628** | **0.8403** |
| $(2,*)$ | **0.5618** | **0.8623** | 0.5618 | 0.8399 |

(d) Freesound MAP/MAROC

**Figure 2**: Performance metrics for text-based retrieval and annotation of sounds, respectively. Data is averaged across $n = 50$ trials with half the tagging data missing. Curves are labeled according to the order of path lengths used in sorting results, where $*$ denotes all shortest paths. $Limit$ is the absolute best performance possible with the dataset.

cross-validation, where the set of sounds and their associated tags are split into several (e.g. 10) random non-overlapping subsets, the classifiers are trained with only one subset, and the remaining sounds are used as queries to test the performance of the trained classifiers. With a sufficiently large training dataset, performance results should converge to give a picture of the expected performance in a production setting.

In [19], [20], and [13], this technique was employed for shortest-path retrieval. For the cases of retrieval and annotation using sounds and tags not present in the training data (thereby testing the usefulness of both acoustic and semantic similarity), sounds and tags were split into 2 and 5 subsets, respectively, each combination thereof (one of $2 \times 5 = 10$) being used to build the network. For annotation, out-of-network sound queries were independently introduced to the network by computing their similarity to all other sounds in the network, and out-of-network tags were connected only to the in-network tags. Query performance was tested by querying each out-of-network sound against out-of-network tags. For retrieval, tag queries were connected independently to other tags, and out-of-network sounds were connected only to in-network sounds.

However, this method of cross-validation may not be entirely appropriate for shortest-path retrieval, as there is no distinct training phase (it is non-parametric). Rather than having only sounds and tags as training data, acoustic and semantic similarities *between* training instances must be considered. For this reason, we have chosen to implement a different evaluation strategy to compare techniques. In this strategy, we simulate a database where the set of sounds and tags are complete (there are no cross-validation splits), but only a random subset of the user tagging data is available. Specifically, for each association between a sound and a tag (for which there may be many for a single sound-tag pair in the *Soundwalks* dataset), we remove it with 50% probability. For annotation, every sound is used to query the entire set of tags, and for retrieval, every tag is used to query the entire set of sounds. Relevance results are then averaged over each query and over 50-100 trials with different tagging data. This simulation is perhaps more appropriate than the networks built for cross-validation, as we can examine how using shortest-path retrieval can help make up for sparse tagging data, which is oftentimes present in online tagging systems.

#### *4.2.2 Performance measures*

Each query returns an ordered list of nodes (tags for annotation and sounds for retrieval), sorted by path length in ascending order. An item in this list is said to be *relevant* if it is connected to the query at least once in the original user tagging data. Using this list of relevance for each item returned, we can compute mean *precision*, the percentage of items returned that are relevant as more items are returned, and mean *recall*, the percentage of all relevant items that have been returned. Plotting precision as a function of recall is a useful way of comparing different schemes. Additionally, one can compute summary statistics including *mean average precision* (MAP), the mean of precision values at the points where each relevant item is returned, and *mean area under the receiver operator characteristic* (MAROC), the integral of the curve produced by plotting the ratio of true positives versus false positives.

### 4.3 Depth-ordered retrieval

In Figure 2, we examine the effects of a) using shortest paths versus retrieving items based only on their tags (as most tag-based search strategies do) and b) using differ-

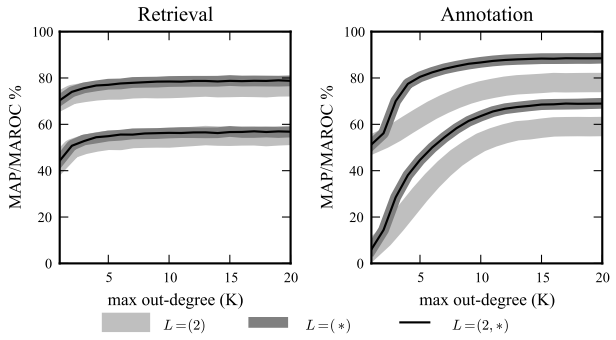



**Figure 3**: Effects of pruning outbound edges to nodes' K nearest neighbors using the *Soundwalks* dataset. Both MAP (lower) and MAROC (upper) values are plotted as a function of K and averaged across $n = 100$ trials with half the tagging data missing.

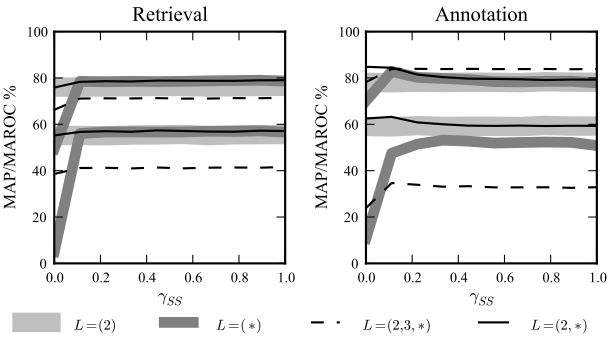



**Figure 4**: Effects of varying $\gamma_{SS}$, the global weight multiplier for sound-to-sound edges using the *Soundwalks* dataset. Both MAP (lower) and MAROC (upper) values are plotted as a function of $\gamma_{SS}$ and averaged over $n = 50$ trials with half tagging data missing.

ent depth ordering strategies. $L = (2)$ corresponds to the case where only direct tag-to-tag links are used for retrieval (the baseline case, given no acoustic similarity information), $L = (*)$ represents the case of ranking results based on the lengths of their shortest paths, and $L = (2, *)$ and $L = (2, 3, *)$ represent the cases of returning minimum 2- and 3-node paths before resorting to shortest paths. Results are shown for both the *Freesound* and *Soundwalks* datasets. $Limit$ represents the theoretical upper limit on performance imposed by the dataset, where the ground truth user tagging data itself is used to order results, analagous to $UpperBnd$ from [16]. Acoustic links were included for the *Soundwalks* dataset but not the *Freesound* dataset, in order to study the effects of using shortest-path retrieval as a drop-in method in an existing system.

From these plots, we can see that, in some cases, as in annotation on the *Soundwalks* dataset (Figure 2b), using shortest paths performs worse than the baseline case, likely because known sound-to-tag links are being circumvented in favor of paths that use acoustic similarity. However, $L = (*)$ seems to perform marginally better than $L = (2)$ for the case of retrieval. To account for this difference, we can see that prioritizing direct links, as in $L = (2, *)$, performs best. $L = (2, 3, *)$ is a special case, as it produces higher MAP/MAROC, corresponding to its better performance in the last 75% of results, but it initially performs quite a bit poorer at annotation, which may be undesirable (if, say, we were to annotate with only those tags that score highest).

For the *Freesound* dataset, for which we provided no sound-to-sound links, we can see that the $L = (*)$, and optionally $L = (2, *)$, methods can assist in ordering the last half of results. This improvement is likely because, for annotation (and analagously for text-based retrieval), a sound can be annotated with additional tags from those sounds it shares a few tags with. Of course, in some use cases, this increase in performance may not be worth the extra query time. Note that $L = (2, 3, *)$ would behave the same as $L = (2, *)$ in this case, as no sound-to-tag paths with an odd number of nodes exist in a network containing no sound-to-sound edges.

### 4.4 Pruning

To test the effects of limiting search to nodes' K nearest neighbors, we first constructed a network as described in 4.2.1 using the *Soundwalks* dataset, with sound-to-sound and sound-to-tag links, but with half the tagging data missing. For $K \in \{1, 2, ..., 20\}$, we then annotated with each sound and retrieved with each tag, testing relevance against the original tagging data. For each value of $K$, we averaged performance metrics over 50 trials for a total of 1000 trials per query type. As shown in Figure 3, it is only when $K < 10$ that significant losses in MAP/MAROC can be seen, suggesting that edge pruning can drastically improve query time without having significant effects on performance, as $10 \ll |E|$.

### 4.5 Weighting edge classes

Figure 4 demonstrates that, for the *Soundwalks* dataset, there is a clear shift in performance at $\gamma_{SS} \approx 0.2$. For $\gamma_{SS} < 0.2$, acoustic weights are used as the primary source of similarity information at the expense of known tagging data. For the case of annotation, there appears to be a slight increase in MAROC for $L = (*)$ at this point. For $L = (2, *)$, there is a slight increase in performance for $\gamma_{SS} < 0.2$, which suggests that the system performs slightly better when tagging data is used for direct links, but acoustic similarity, rather than cooccurrence of tags, is primarily relied on when no direct, 2-node links exist.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we have experimented with several modifications of a shortest-path retrieval algorithm, where acoustic similarities between sounds are used in conjunction with user tagging data for the purpose of annotation and text-based retrieval. Specifically, we have demonstrated that:

1. giving priority to direct, 2-node paths before resorting to shortest path lengths can greatly improve annotation and retrieval accuracy,

2. pruning edges searched to nodes' $K$ nearest neighbors can reduce query complexity from $O(|V|^2)$ to $O(|V|\log|V|)$ for values of $K$ as low as 10, and

3. relative weighting between edge classes (sound-to-sound versus sound-to-tag) influences retrieval results only slightly but indicate when certain types of similarity information are best used.

While the query time of this approach will likely be slower than similar parametric classification approaches for any database where the number of desired classifiers (tags) is much less than the number of sounds, this approach can still be very useful for smaller datasets, where acoustic similarity between sounds and co-occurrence of tags can help make up for sparse tagging data, as shown in the results for the *Freesound* dataset.

In [20], a method where spectral clustering is used to create cluster nodes that reduce the number of sound-to-sound edges is discussed. In addition to actually improving query accuracy, query complexity is greatly reduced. Combined with the methods of depth-ordered search and pruning we have introduced in this paper, pre-processing sound-to-sound edges in this way could achieve time complexity as low as $O(\log|V|)$ during queries.

Additionally, [13] and [19] discuss the effects of the presence of tag-to-tag edges. While it was shown in [19] that tag-to-tag edges between in-network tags tend to hinder performance, using different edge class weights ($\gamma_{SS}$ and $\gamma_{TT}$) and depth-ordered search could create a situation where tag-to-tag edges can improve results.

## 6. ACKNOWLEDGEMENTS

Special thanks are given to Gordon Wichern and our reviewers for many helpful comments and the contributors to and developers of *Freesound.org* for providing a wonderful source of reusable, tagged sounds. This material is based upon work supported by the National Science Foundation under grant NSF IGERT DGE-05-04647 and NSF Net Centric I/UCRC - ASU SenSIP Site Award 1035086.

## 7. REFERENCES

[1] Luke Barrington, Mehrdad Yazdani, Douglas Turnbull, and Gert Lanckriet. Combining feature kernels for semantic music retrieval. In *International Conference on Music Information Retrieval*, pages 614–619, 2008.

[2] P. Cano and M. Koppenberger. Automatic sound annotation. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 391–400, 2004.

[3] 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 *The 116th AES Convention*, Berlin, Germany, 2004.

[4] Fabio Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11:453–482, 1997.

[5] S. Essid, G. Richard, and B. David. Inferring efficient hierarchical taxonomies for music information retrieval tasks: Application to musical instruments. In *International Conference on Music Information Retrieval*, pages 324–328, 2005.

[6] M. Goto and K. Hirata. Recent studies on music information processing. *Acoustical Science and Technology*, 25(6), 24.

[7] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

[8] B. H. Huang and L. R. Rabiner. A probabilistic distance measure for hidden markov models. *AT&T Technical Journal*, 64(2):1251–1270, 1985.

[9] S. Kim, S. Narayanan, and S. Sundaram. Acoustic topic model for audio information retrieval. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 37–40, New Paltz, NY, 2009.

[10] T. Li and M. Ogihara. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 239–240, Baltimore, MD, 2003.

[11] O. Celma M. Sordo, C. Laurier. Annotating music collections: How content-based similarity helps to propagate labels. In *International Conference on Music Information Retrieval*, Vienna, Austra, 2007.

[12] E. Martinez, O. Celma, M. Sordo, B. de Jong, and X. Serra. Extending the folksonomies of freesound.org using content-based audio analysis. In *SMC*, Porto, Portugal, 2009.

[13] B. Mechtley, G. Wichern, H. Thornburg, and A. S. Spanias. Combining semantic, social, and acoustic similarity for retrieval of environmental sounds. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, 2010.

[14] T. Pederson, S. Patwardhan, and J. Michelizzi. Wordnet::similarity: measuring the relatedness of concepts. In *Innovative Applications of Artificial Intelligence Conference*, pages 1024–1025, Cambridge, Massachusetts, 2004. AAAI Press.

[15] M. Slaney. Semantic-audio retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages IV–1408–IV–1411, 2002.

[16] 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*, 16(2):467–476, February 2008.

[17] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[18] B. Whitman and D. Ellis. Automatic record reviews. In *International Conference on Music Information Retrieval*, pages 470–477, 2004.

[19] G. Wichern, B. Mechtley, A. Fink, H. Thornburg, and A. Spanias. An ontological framework for retrieving environmental sounds using semantics and acoustic content. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[20] G. Wichern, H. Thornburg, and A. Spanias. Unifying semantic and content-based approaches for retrieval of environmental sounds. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 13–16, New Paltz, NY, 2009.

[21] 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*, 18(3):688–707, 2010.