COMBINING SEMANTIC, SOCIAL, AND ACOUSTIC SIMILARITY FOR RETRIEVAL OF ENVIRONMENTAL SOUNDS

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

Arizona State University
School of Arts, Media, and Engineering and SenSIP Center
Tempe, AZ 85282
{bmechtley, gordon.wichern, harvey.thornburg, spanias}@asu.edu

ABSTRACT

Recent work in audio information retrieval has demonstrated the effectiveness of combining semantic information, such as descriptive, tags with acoustic content. However, these methods largely ignore the possibility of tag queries that do not yet exist in the database and the possibility of similar terms. In this work, we propose a network structure integrating similarity between semantic tags, content-based similarity between environmental audio recordings, and the collective sound descriptions provided by a user community. We then demonstrate the effectiveness of our approach by comparing the use of existing similarity measures for incorporating new vocabulary into an audio annotation and retrieval system.

Index Terms— acoustic signal analysis, database query processing, semantic networks, multimedia databases

1. INTRODUCTION

There has been much recent work within the audio signal processing community in combining audio content analysis with semantic information such as user-provided tags or descriptions (which we will also refer to as concepts). These methods include connecting audio features with semantic words through hierarchical clusters [1], using audio features to train classifiers for each word in a vocabulary [2, 3], and using a WordNet [4] based taxonomy to automatically describe sounds based on the descriptions of their nearest neighbors in an audio feature space [5].

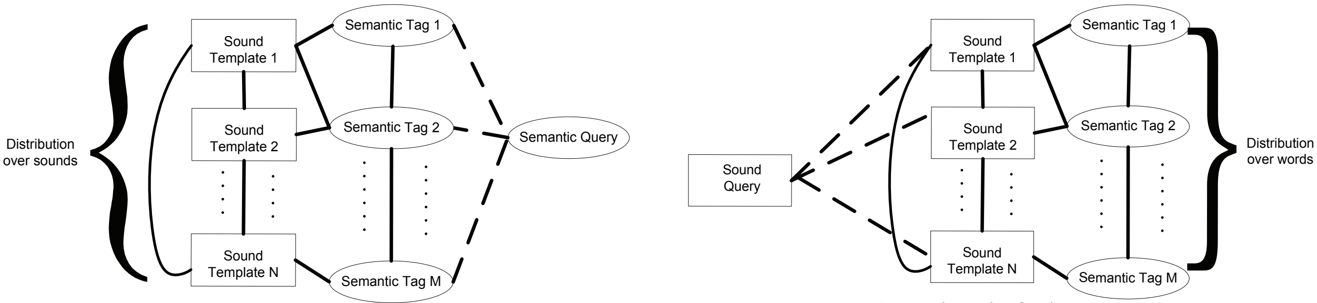
However, almost no techniques take into consideration semantic queries where the words used as queries have not yet been used to describe a sound. In a social network or other interactive application, this can correspond to the time before a tag has been applied enough times (most likely by the same person who introduced it) to significantly influence retrieval results. Additionally, these techniques do not leverage semantic similarity between concepts. For example, if the words “purr” and “meow” are independently applied to separate sounds, the retrieval system has no way of knowing that these sounds may have been emitted from the same physical source even though they are widely separated in the acoustic feature space.

Fortunately, several techniques have been developed for measuring semantic similarity between concepts. These techniques include using dictionaries to compare definitions of terms, using thesauri to compare overlaps in synonymous words, and using hierarchical ontologies such as WordNet to compute path-related distances [6].

In the present work, we extend our previous approach for unifying semantic and content-based approaches to audio information retrieval [7] by incorporating quantitative measures of semantic similarity for the concepts used to describe sounds. Inspired by the success in organizing lexical databases such as WordNet, we utilize an ontological framework where all sounds and tags in the database each represent a node in a network. Specifically, sounds are linked to other sounds based on a measure of perceptual similarity, semantic tags are linked to each other through an external similarity computed with WordNet, and semantic tags and sounds are connected based on descriptions provided by a user community. Obtaining the connections between semantic tags and sounds requires user input and is often achieved by incorporating a social or game element into the tag collection process [8]. Once all links in the ontological framework are obtained, it forms a flexible system that allows for content-based retrieval, automatic tagging of unlabeled audio, and text-based search as discussed in [7].

By including quantitative measures of semantic similarity our system allows for two new audio information retrieval tasks: a) *text-based retrieval with foreign concepts* and b) *annotation with foreign concepts*. In (a), one can retrieve sounds based on semantic tags that do not yet exist in the database. Similarly, with (b), one can use a sound query to obtain a distribution over a set of concepts that do not exist in the database. Experimental results on a database of sounds with community semantic descriptions demonstrate the ability of

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



(a) Text-based retrieval with foreign concepts.

(b) Annotation with foreign concepts.

Fig. 1. Semantic and audio queries using concepts not present in the database. Dashed lines indicate links added at query time.

the proposed approach for the retrieval tasks mentioned above while also comparing the performance of different semantic similarity measures.

2. RETRIEVAL NETWORK

The application we have described requires the storage and retrieval of acoustic, semantic, and social information. To perform retrieval in such a way that is consistent, a unifying structure is needed that represents all types of knowledge present in the database. To this end, we have developed [7] an undirected network that can be used to perform probabilistic queries over any subset of nodes, i.e. sounds or concepts. The network (Figure 1) consists of a weighted, undirected graph of nodes, each of which can be sounds ($s_{1:N}$) or concepts ($c_{1:M}$). Possible link types include sound-to-sound, concept-to-concept, and sound-to-concept links that have weights $W(i,j)$.

2.1. Acoustic information: sound-to-sound links

Sound-to-sound weights can be computed by comparing the acoustic content of each sound. This process begins with acoustic feature extraction, where six features are calculated using overlapping 40ms Hamming windows hopped every 20ms. The feature trajectory for a sound file is given by $Y_{1:T}^{(1:P)}$ where $Y_t^{(i)}$ is the ith feature value at frame t.

The six features we use include *loudness*, the dB-scaled RMS level over time; temporal sparsity, the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one-second interval; spectral sparsity, the ratio of ℓ^∞ and ℓ^1 norms calculated over the short-time Fourier Transform (STFT) magnitude spectrum; bark-weighted *spectral centroid*, a measure of the average frequency content of a sound at any point in time; transient index, the ℓ^2 norm of the difference of Mel frequency cepstral coefficients (MFCC's) between consecutive frames; and harmonicity, a probabilistic measure of whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure. For more details on how these features are calculated, see [9]. This feature set was developed for a broad range of environmental sounds rather than any specific class of sounds (e.g. speech or music) and with a focus on *ecological validity*, such that the features would best relate to qualities of the human auditory mechanism by using perceptual scalings of spectral content [10, 9].

To compare sounds, [11] describes a method of estimating $L(s_i, s_j) = log(P(Y_{1:T}^{(1:P)}(s_i)|\lambda^{(1:P)}(s_j)))$, the log-likelihood that the feature trajectory of sound s_i was generated by the hidden Markov Model (HMM) $\lambda^{(1:P)}(s_j)$ built to approximate the simple feature trends of sound s_j.

For retrieval, however, it is helpful to have a semi-metric, $D(s_i, s_j)$, between sounds that is *symmetric* ($D(s_i, s_j) = D(s_j, s_i)$) and *nonnegative* ($D(s_i, s_j) \geq 0$). In [7, 12], a semi-metric that holds these properties is given as:

$$W(s_i, s_j) = L(s_i, s_i) + L(s_j, s_j) - L(s_i, s_j) - L(s_j, s_i). \tag{1}$$

2.2. Semantic information: concept-to-concept links

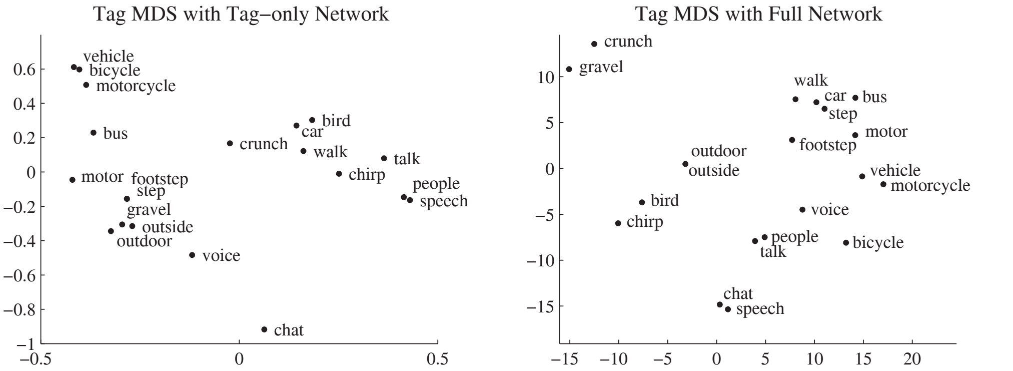
For concepts, we obtain a fully connected concept-to-concept similarity matrix using a similarity metric from the WordNet::Similarity library [13]. Specifically, we have used the `jcn`, `lin`, `lch`, `res`, and `vector` metrics. The `vector` metric computes the co-occurrence of two concepts within the collections of words used to describe other concepts (their *glosses*) [13] and `lch` is based on the shortest-path distance between two concepts in a taxonomy, while `jcn`, `lin`, and `res` are all based on the information content of the terms and their ancestors within a corpus (for a full review, see [6, 13].)

For consistency in comparing different metrics, we compute the network weights from the following normalization of similarities, $S(c_i, c_j)$:

$$W(c_i, c_j) = -\log\left[\frac{S(c_i, c_j)}{\max_{i,j} S(c_i, c_j)}\right]. \tag{2}$$

2.3. Social information: sound-to-concept links

Letting V be an $M \times N$ votes matrix, where V_{ij} is equal to the number of users who have tagged sound s_i with concept c_j



(a) network containing only semantic weights

(b) network containing semantic, social, and acoustic weights

Fig. 2. MDS of tags for networks with and without non-semantic information using the `vector` semantic similarity metric.

divided by the total number of users who have tagged sound s_i, we can compute the joint probability between s_i and c_j as $P(s_i, c_j) = V_{ij} / \sum_{k,l} V_{kl}$. We can then define the weight as $W(s_i, c_j) = -\log P(s_i, c_j)$, which is a simple approximation of the Kullback-Leibler weight learning technique proposed in [7].

In our application, the votes matrix is obtained through tagging, but in the future it could be expanded to reflect more implicit community activity, such as discussions, rankings, or page navigation on a website.

2.4. Retrieval

Given a subset of nodes, $\mathcal{A}$ and query node, q, a posterior distribution from the network can be calculated as follows:

$$P(a \in \mathcal{A}|q) = \frac{e^{-d^*(q,a)}}{\sum_{b \in \mathcal{A}} e^{-d^*(q,b)}}, \qquad (3)$$

where $d^*(q, a)$ is the shortest-path distance in the network between nodes q and a, which can be efficiently computed using Dijkstra's algorithm [14]. Note that, in the case of a query node that does not yet exist in the database, such as a new sound or concept, the distances between the query node and all other nodes of its type can be computed on demand.

3. RESULTS

Our experiment consists of 178 sounds recorded from 7 separate field recording sessions, lasting anywhere from 10 to 30 minutes each and sampled at 44.1KHz. Each session was recorded continuously and then hand-segmented by the authors into segments lasting between 2-60s. 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 100 most popular tags, leaving 12 out due to part-of-speech incompatibility with certain semantic similarity measures.

Retrieval performance

The retrieval tasks tested are those summarized in Figure 1 i.e., a) *text-based retrieval with foreign concepts*, where a tag not present in the database is used to retrieve a ranked list of sounds, and b) *annotation with foreign concepts*, where a sound is annotated with tags not in the database. To test these tasks, we have used a five-fold cross-validating approach, partitioning tags (or sounds, respectively) into 5 random non-overlapping subsets used for queries and using the remaining 80% to build the network.

To compute two common measures of information retrieval performance, mean average precision (MAP) and mean area under the receiver operation characteristic curve (MAROC), we first rank the number of retrieved nodes (sounds for task (a), tags for task (b)) in order of decreasing probability. A precision curve can then be plotted moving down this list, computing the running percentage of results that are relevant to the query. MAP can then be calculated as the mean of this curve averaged over queries and trials. A ROC curve is computed similarly, but is defined as a the running ratio of true-positive to false-negative retrieval results. MAROC is then the integral of this curve averaged over queries and trials. Both MAP and MAROC are then averaged over all five cross-validation runs to give the final results.

Table 1. Performance of retrieval tasks using different measure of semantic similarity.

	text-based retrieval		annotation	
Measure	MAP	MAROC	MAP	MAROC
`jcn`	**0.2035**	0.6912	**0.7947**	**0.9045**
`lin`	0.1801	0.7211	0.5669	0.8067
`lch`	0.1831	0.6823	0.5560	0.7890
`res`	0.1941	**0.7212**	0.6023	0.8371
`vector`	0.172	0.6855	0.5987	0.8305

Table 1 lists MAP and MAROC values for both tasks using a variety of different semantic similarity measures. The Jiang and Conrath measure [15] seems to perform best, which is consistent with findings for other applications, such as determining word senses in text [16]. Precision values are much lower for both tasks than corresponding tasks in [7, 2, 3], however, it should be noted that the use of foreign terms makes the tasks inherently more difficult. MAROC is quite high in all cases, but this could be misleading, as using the top 100 most popular tags may inflate recall significantly. However, from this data we are able to form a clear idea of which semantic similarity measures work best for audio annotation and retrieval. Future studies will need to include all or randomly-selected user-submitted tags for more accurate performance results. It would also be prudent to test the system across more semantic similarity measures.

Multidimensional scaling

One way to visualize how social, semantic, and acoustic information are combined is to organize their nodes using multidimensional scaling [17]. In MDS a distance matrix (in this case obtained by finding shortest-path distances between all node pairs) is used to organize points into a suitably low dimensionality in such a way that retains their distances.

Figures 2(a) and 2(b) display the two-dimensional MDS for a subset of hand-picked tags. In Figure 2(a) the distance matrix is calculated from a network containing only tag nodes, i.e., only semantic information, while Figure 2(b) contains both sound and tag nodes and the links between them, i.e., acoustic, social, and semantic information. The differences between the absolute scales of the axes in Figures 2(a) and 2(b) result from the different distance matrices, furthermore, in a retrieval/annotation context when a ranked list is returned we are only concerned with comparing relative tag positions. From Figure 2(a), we can see that natural clusters form from the semantic information, such as $\{$*vehicle*, *bicycle*, *motorcycle*, *bus*$\}$. Similarly, synonyms such as *outside*/*outdoor* are near each other. However, some concepts we would think as similar do not cluster, such as *chat*/*speech*.

By including social and acoustic information in the network in Figure 2(b), the concepts organize into clusters that are informed by which concepts sound alike. For example, *chat*/*speech* are now quite near other. Similar new clusterings can be seen between word pairs such as *gravel*/*crunch* and *bird*/*chirp*. Some clusterings are more vague in the reorganization, such as that which formerly grouped all vehicles, as we are now also capturing information of what concepts typically are heard together or in similar circumstances.

4. CONCLUSIONS

Previous methods of text-based retrieval, annotation, and query-by-example in acoustic information retrieval do not leverage semantic similarity between concepts tagged to sounds. In some cases, such as querying or annotating with tags that do not yet exist in the database, this can be prohibitive. Using existing techniques for measuring semantic similarity, such as through measures on the WordNet lexical database, can assist in these cases. Similarly, using a flexible retrieval system such as our demonstrated ontological framework can allow a retrieval system to be augmented quite easily for different applications.

5. REFERENCES

[1] M. Slaney, “Semantic-audio retrieval,” in *IEEE ICASSP*, Orlando, FL, 2002.

[2] D. Turnbull, L. Barrington, and G. Lanckriet, “Semantic annotation and retrieval of music and sound effects,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, February 2008.

[3] G. Chechik, E. Le, M. Rehn, S. Bengio, and D. Lyon, “Large-scale content-based audio retrieval from text queries,” in *ACM MIR*, Vancouver, BC, 2008.

[4] Christiane Fellbaum, *WordNet: an electronic lexical database*, MIT Press, Cambridge, MA, 1998.

[5] 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 *Proc. 116th AES Convention*, Berlin, Germany, 2004.

[6] A. Budanitsky and G. Hirst, “Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures.,” in *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittburgh, PA, 2001.

[7] G. Wichern, H. Thornburg, and A. Spanias, “Unifying semantic and content-based approaches for retrieval of environmental sounds,” in *IEEE WASPAA*, New Paltz, NY, October 2009.

[8] M. Mandel and D. P. W. Ellis, “A web-based game for collecting music metadata,” *J. New Music Research*, vol. 37, no. 2, pp. 151–165, 2008.

[9] G. Wichern, H. Thornburg, B. Mechtley, A. Fink, A. Spanias, and K. Tu, “Robust multi-feature segmentation and indexing for natural sound environments,” in *IEEE CBMI*, Bordeaux, France, July 2007.

[10] A. S. Bregman, *Auditory Scene Analysis*, The MIT Press, Cambridge, MA, 1990.

[11] G. Wichern, J. Xue, H. Thornburg, and A. Spanias, “Distortion-aware query-by-example of environmental sounds,” in *IEEE WASPAA*, New Paltz, NY, October 2007.

[12] J. Xue, G. Wichern, H. Thornburg, and A. Spanias, “Fast query-by-example of environmental sounds via robust and efficient cluster-based indexing,” in *IEEE ICASSP*, Las Vegas, NV, April 2008.

[13] T. Pederson, S. Patwardhan, and J. Michelizzi, “Wordnet::similarity - measuring the relatedness of concepts,” in *AAAI-04*, Cambridge, MA, 2004, pp. 1024–1025, AAAI Press.

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

[15] J. Jiang and D. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” in *International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan, 1997, pp. 19–33.

[16] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll, “Finding predominant word senses in untagged text,” in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004, pp. 280–287.

[17] J.B. Kruskal and M. Wish, *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA, 1978.